Applying Item Response Theory Models to the Survey of Adult Skills (PIAAC)

Geremew, Netsanet Meseganaw
Masteruppsats 30 högskolepoäng, vt 2014

Supervisor: Dan Hedlin
Abstract

The analysis of test generated data or survey response often utilizes means, total scores or latent scores from factor analysis as outcome variables in a regression based procedure. These scoring methods however, do not take into account the characteristics of the items in the test or survey. Item response theory (IRT) is different in this sense; it models the relationship between a respondents trait level (ability) and the pattern of items response. In this thesis we will illustrate the use of unidimensional item response theory to analyse polytomous response data. This study also demonstrates the use of IRT model for differential item functioning (DIF) by allowing for group effects on item response functioning (IRFS). Application of IRT is illustrated through a comparative educational assessment survey called Program for the International Assessment of Adult Competencies (PIAAC). The current paper provides a basic overview of item response theory analysis for assessing a latent factor structure. The objective is to evaluate an underlying variable related to an issue problem solving ability of adults and utilize a graded response model in IRT. Discussion includes methodology of IRT, analysis, findings, and implications of IRT for educational research. The result obtained show that IRT is useful tool for both test theory and test development. Nevertheless the assumption of unidimensional and local independence of items makes the model limited in several types of test.

Key words: Item Response Theory (IRT), Differential Item Functioning (DIF) and Graded Response Model (GRM).
Acknowledgements

I would like to give my special gratitude and deepest respects to my supervisors Professor Dan Hedlin from the Statistics Institute of Stockholm University and I am thankful to his patience, guidance and support throughout the process of writing this master thesis.

I am also very grateful to my family and my friends for their warmest encouragement.
# CONTENTS

1 Introduction .......................................................................................................................... 4
2 General Overview of IRT ...................................................................................................... 6
   2.1 General Form of IRT models .......................................................................................... 7
       2.1.1 The two parametric logistic model (2PLM) ......................................................... 8
       2.1.2 The graded response model (GRM) ..................................................................... 10
           2.1.2.1 Model formulation and analysis .................................................................. 12
   2.2 Model Assumptions ........................................................................................................ 13
3 Program for the International Assessment of Adult Competencies (PIAAC) .................... 16
   3.1 Data Description ............................................................................................................ 16
   3.2 Assessment of Model Assumption .............................................................................. 19
4 Data Analysis and Result .................................................................................................... 24
   4.1 Goodness of fit tests ..................................................................................................... 24
   4.2 Item Parameter Estimation ............................................................................................ 25
   4.3 Test characteristic curve ............................................................................................... 26
   4.4 Test information Curve ................................................................................................. 29
   4.5 Test standard error curve ............................................................................................. 32
   4.6 Reliability Analysis ....................................................................................................... 33
   4.7 Differential Item Functioning (DIF) ........................................................................... 36
       4.7.1 Analysis for comparison of countries ................................................................ 36
5 Discussion ............................................................................................................................ 40

References ............................................................................................................................... 42
Appendix A .............................................................................................................................. 44
   Table and Figures ................................................................................................................ 44
Appendix B .............................................................................................................................. 51
   Applied statistical methods ............................................................................................... 51
List of tables

Table 1. Participating countries in PIAAC and sample size..............................................16
Table 2. Four items used for the analysis...........................................................................18
Table 3. Top US and bottom Sweden LD X² Statistics .........................................................23
Table 4. Top US and Sweden Factor Loadings.......................................................................23
Table 5. Top US and bottom Sweden s – X² Item Level Diagnostic Statistics .................25
Table 6. Top US and bottom Sweden Graded Model Item Parameter Estimates.................28
Table 7. Top US bottom Sweden Reliability analysis of the problem solving test.............35
Table 8. Top US and bottom Sweden, DIF Statistics for Graded Items .........................39

List of Appendix A table

Table A.1. Table 5 GRM without covariate Likelihood-based Values and Goodness of Fit Statistics US ........................................................................................................44
Table A.2. GRM with covariate Likelihood-based Values and Goodness of Fit Statistics US...44
Table A.3. GRM without covariate Likelihood-based............................................................45
Table A.4. GRM with covariate Likelihood-based.................................................................45
Table A.5. Graded Model Item Parameter Estimates for Group 1, US...............................46
Table A.6. Graded Model Item Parameter Estimates for Group 2, US...............................46
Table A.7. Graded Model Item Parameter Estimates for Group 1, Sweden.......................47
Table A.8. Graded Model Item Parameter Estimates for Group 2, Sweden.......................47

List of Appendix B table

Table B.1. Latent variable models ........................................................................................51

List of figures

Figure 1. Represents the location of the items and individualas on the same continuum of the latent trait Person’s location is the measure of the person’s latent trait and item’s location. .......7
Figure 2. 2PLM trace lines with different location (b) parameters.......................................9
Figure 3. 2PLM trace line. This function indicates the relationship between the latent trait (θ) and the probability of choosing the keyed alternative.........................................................9
Figure 4. 2PLM trace lines with different slope (a) parameters..........................................10
Figure 5. GRM trace lines the upper figure illustrate the meaning of the b values for 5 option item the lower panel illustrate the model based probability of selecting each response option ....11
Figure 6. Top Sweden and Bottom US. Bar plot displays the raw score of female and male adults answer for the question problem solving domain by country. Error! Bookmark not defined.

Figure 7. Top Sweden and Bottom US scree plot of the eigen values, the scree clearly begins to flatten at the second eigen value. ...............................................................21

Figure 8. Top US and bottom Sweden Test characteristic curves ........................................27

Figure 9. Top top US and bottom Sweden, Test information Curves ................................31

Figure 10. Top US bottom Sweden standard error curves ..................................................32

Figure 11. Average for problem solving for US and Sweden ..............................................39

Figure 12. Sweden top group 1 and bottom group 2 total information curves ................37

List of Appendix A Figure

FigureA. 1. Graphical examples of software IRTPRO 2.1 ..................................................49

FigureA. 2. US top group 1 and bottom group 2 test characteristic curves .........................49

FigureA. 3. Sweden top group 1 and bottom group 2 test characteristic curves ..................50

Figure A. 4. US top group 1 and bottom group 2 total information curves .......................50
1 Introduction

Item Response Theory (IRT) models are commonly used to model the latent traits associated with a set of items or questions in a test or survey. In education, testing is an inherent part of the curriculum as an assessment tool to evaluate students’ subject matter proficiency and skill development. Apart from viewing the total score as an indicator of performance one may wish to understand whether the testing instrument is adequately designed to measure particular aspects of the knowledge and skills of respondents. IRT attempts simultaneously to examine the appropriateness of the questions in terms of measuring what they are designed to measure and the proficiency of the respondents.

Other applications of IRT besides assessment include: item functioning, designed for group comparisons based on external or non-construct-related factors IRT models describe the interactions of persons and test items (Reckase, 2008). Hence, IRT is a general framework for specifying mathematical functions that characterize the relationship between a person’s ability or trait as measured by an instrument and the person’s responses to the separate items in the instrument (DeMars, 2010). In educational testing, IRT offers an alternative to classical test theory, which depends on total scores or number correct as outcome variables. IRT models have become a popular framework in many fields including psychology, nursing, and public health. But IRT-based analysis remains scarce in the social sciences and in the educational research literature.

Currently IRT is finding widespread application in the engineering of large-scale assessments as well as on a smaller scale in sociological and psychological assessments. Classical test theory (CTT) was the dominant approach until 1953 when Frederic Lord published his doctoral dissertation on Latent Trait Theory. While CTT models test outcomes based on the linear relationship between true and observed score (Observed score = True Score + Error), IRT models the probability of a response pattern of an examinee as a function of the person’s ability and the characteristics of the items in a test or survey.

The focus of this study is the application of unidimensional IRT (UIRT) on Problem solving in technology-rich environments (PSTRE) ability of adults in two countries Sweden and US. The data for this analysis come from the Program for the International Assessment of Adult Competencies (PIAAC) 2012.

The softwares used in this work are IRT PRO for windows 2.1, Database Analyzer (IDB Analyzer, 2013) and SPSS (Statistical Package for the Social Sciences). IRT PRO for windows 2.1 is developed by li Cai, David Thissen and Stephen du Toit. This product has replaced the four programs; Bilog-MG, Multilog, parscale, testfact which overlap in functionality more about IRT PRO for windows is explained in the appendix A part. The IEA International Database Analyzer (IDB Analyzer, 2013) is an application developed by the IEA Data Processing and Research Center (IEA DPC) in Hamburg, Germany. The IDB Analyzer can be used to combine and analyze data from IEA’s large-scale assessments.

The remainder of the study is organized as follows. Section 2 presents an overview of the IRT models in the context of an educational testing situation in which the objective is to assess both individual ability and question difficulty. Section 3 introduces the survey of adult skill called the Program for the International Assessment of Adult Competencies (PIAAC). The specific model used, and issues involved with estimation and model assessment (goodness of fit) are discussed. Results of the analysis are presented in Section 4. Section 5 contains discussion remarks. More technical details on the formulation of IRT models and parameter estimation are presented in Appendices A and B.
2 General Overview of IRT

IRT modeling has a long history and extensive literature. In this section, we provide a brief overview of some popular IRT models and their assumptions. The discussion of IRT can be found in Baker F.B., and Kim, S.H. (2004) and Bock (1997). IRT is an approach to modern educational and psychological measurement which addresses the measurement of a hypothetical latent construct such as ability or attitude. These latent traits cannot be measured directly on individuals and must be quantified via responses to items or questions in a test or survey.

IRT methods are commonly used to obtain latent scores for individual respondents on qualities such as trait, ability, proficiency, or attitude in a test or survey. IRT is perhaps most easily understood in terms of the latent trait ability in a testing situation. In fact, the first applications of IRT were in educational testing. The IRT scoring process takes into account the respondent’s latent variable and the item’s characteristics such as difficulty and discrimination.

Applications of IRT are used in many fields such as psychometrics, educational sciences, sociology, health professional fields and computer adaptive testing (CAT). Additionally, IRT can be used in test or instrument development because the IRT models utilize information about the item’s characteristics to evaluate and renew an instrument.

IRT models, in contrast to classical test theory (CTT), do not rely on sums or number of correct scores to evaluate a person’s performance. Nor do they assume equal contribution of the items (questions) to the overall scores. Since items vary in their difficulty and persons vary in their trait level, this method may result in a more accurate assessment of respondents’ latent traits because respondents with the same sum score may differ in their trait measurement.

IRT methods also use the same metric to measure the latent variable and the items’ difficulty levels, thereby facilitating comparison and estimation of parameters. Figure 1 displays the placement of item location and person trait level on the same scale. If the difference between a person’s location (ability) and an item’s location (difficulty level) is positive, the person has a high chance (greater than 50%) of answering that item correctly, that is endorsing that item with a positive score. However, if the difference is negative then the person’s chance of getting an incorrect answer is higher.
Figure 1. The graph represents the location of the items and individuals on the same continuum of the latent trait. A person’s location is the measure of that person’s latent trait and an item’s location is a measure of the item’s difficulty. When the two locations coincide, the person is expected to answer the item correctly or positively with a 50:50 chance. (Source: applying item response theory modeling in educational research by Dai-Trang Le).

2.1 General Form of IRT models

IRT includes a set of models that describe the interactions between a person and the test items. Persons may possess different traits and instruments may be designed to measure more than one trait. In this study we only discuss IRT models that describe one single trait. These models are often referred to as unidimensional IRT (UIRT).

Consider an educational testing situation in which we obtain n individual answer to one question or item. Let j = 1 . . . . . n and i = 1 . . . . I and let $Y_{ij}$ be random variables associated with the response of individual j to item i. These responses may be binary (e.g. correct / incorrect) or may be discrete with a number of categories. Let $\Omega_Y$ denote the set of possible values of the $Y_{ij}$ assumed to be identical for each item in the test. Finally $\theta_j$ denotes the latent trait or ability for individual and $\eta_i$ denote the set of parameters that will be used to model items characteristics. Different IRT models arise from different set of possible response set $\Omega_Y$ and different functional forms assumed to describe the probabilities with which the $Y_{ij}$ assumes those values, namely

$$ p(Y_{ij} = y / \theta_j, \eta_i) = f(y / \theta_j, \eta_i), y \in \Omega_Y $$ (2.1)

The item parameter $\eta_i$ may include two distinct types of parameter, discriminating parameter $\alpha_i$ and a difficulty parameter $b_i$. 
The discriminating parameter $\alpha_i$ in equation 2.1 changes within the change in ability $\theta_j$ the difficulty parameter $b_i$ model how difficult the item is. This will be further illustrated below.

The following section illustrate parametric IRT models

2.1.1 The two parametric logistic model (2PLM).

Using notation similar to These and Steinberg (2010), the two parametric logistic model (2PLM), the probability of endorsing an item under the 2PLM can be expressed as:

$$ P(y_i = 1|\theta) = \frac{1}{1 + \exp(\alpha_i(b_i - \theta))} \quad (2.2) $$

where $P(y_i = 1|\theta)$ is the probability of endorsing the item $i$ in the keyed direction ($y_i = 1$), given an individual’s latent trait score ($\theta$), $\alpha_i$ is the item slope parameter and $b_i$ is the item location parameter. The probability in (2.2) is for one individual’s response to item $i$. The parameter $\theta$ varies with individuals; however, we sometimes suppress the subscripts. The parameter $b_i$ can be interpreted as the difficulty of items, larger values of $b$ are associated with lower proportions of correct responses. This definition assumes that all parameter values for individuals ($\theta$) and items ($\alpha_i, b_i$) are known (i.e., they have been estimated). Figure 2 displays a 2PLM trace line, a graphical visualization of the item properties of a 2PLM item. The location parameter is measured at the point on the trace line at which examinees with the corresponding trait level ($\theta$) has a 50/50 chance of endorsing the keyed response. The item in Figure 2 has a $b$ value of 0. Based on the 2PLM, an individual of average ability (i.e., with a $\theta$ of 0) would have approximately a 50% probability of endorsing this item.

The slope parameter $\alpha$ is called the discrimination parameter of an item. The parameter indicates the extent to which the items discriminate the abilities of the examinees, it represents the steepness of the items response function, and the item specific in the 2PL model to make the model identified. The ability parameter $\theta$ is constrained to have mean of 0 and a
The Rasch model (RM) is often referred to as a one parameter logistic model (1PL); unlike the 2PL model, the slope parameter \( \alpha \)'s are constant across all items. 

![Figure 2](image1.png)

**Figure 2.** 2PLM trace line. This function indicates the relationship between the latent trait (\( \theta \)) and the probability of choosing the keyed alternative.

![Figure 3](image2.png)

**Figure 3.** 2PLM trace lines with different location \( (b) \) parameters.

- Item 2 is more difficult to answer correctly than item 1.
Figure 4. 2PLM trace lines with different slope (a) parameters.

Item 4 is more discriminating than item 3.

2.1.2 The graded response model (GRM)

For analyzing polytomous response data, several have been used. In the following, we review the graded response model (Samejima 1969). The graded response model (GRM) assumes that the cumulative log odds for scoring $k \in \{1, 2, \ldots, K\}$ or higher on item $i$ is a linear function of $\theta$.

$$
\log \left( \frac{Pr(Y_i \geq k | \theta)}{Pr(Y_i < k | \theta)} \right) = \alpha_i (\theta - b_{ik})
$$

(2.3)

where $k$ is response category $\{1 \leq k \leq K\}$, $i$ is an item and $\theta$ is the latent variable such as ability, the discrimination parameter, $\alpha_j$ and the item category step parameter, $b_{ik}$ are ordered by the category index $i$, $b_{i1} < b_{i2} < \ldots < b_{ik-1}$. The GRM is an extension of the 2PLM for items
with polytomous response scales for each item the GRM incorporates a single slope ($\alpha$) parameter and multiple location ($b$) parameters.

![GRM trace lines](image)

![Probability of Response](image)

**Figure 5.** GRM trace lines the upper figure illustrate the meaning of the $b$ values for a 5-category option item. The lower panel illustrate the model based probability of selecting each response option for that same item and $b_{jk}$ are not necessarily ordered.
2.1.2.1 Model formulation and analysis

As previously discussed, an IRT model derives the probability of a response for a particular item in a survey or test as a function of the latent trait $\theta$ and the item parameters. We are interested in the probability of responding in a specific category. In the graded response model, the cumulative probabilities are modeled directly. This is the probability of responding within or above a given category. Then the probability of responding in a specific category is modeled as the difference between two adjacent cumulative probabilities. Let $K$ note the number of response categories of item $i$. For simplicity, we assume that all items have the same $K$ number of unique categories. Then there are $K-1$ thresholds between the response options.

For $i=1, \ldots, I; j=1, \ldots, n; \text{ and } k=1, \ldots, K$. Let $I$ denote the number of items, $n$ the number of persons and $K$ the number of categories which we assume is the same for all items. Let $y_{ijk}$ be response $k$ to item $i$ for person $j$, let $\alpha_i$ represent the discrimination parameter of item $i$ and $b_{ik}$ be the category boundaries or thresholds for category $k$ of item $i$ there for $K-1$ thresholds $b_{iks}$, between the response option. These thresholds are the boundaries between two adjacent cumulative scores, for example $b_{i3}$ is the thresholds between a score of 3 or higher and a score 2 or lower.

The cumulative probabilities have the mathematical representation as in equation.

$$
P(y_{ijk} \geq 1 | \theta_j, b_{i1}, \alpha_i) = 1
$$

$$
P(y_{ijk} \geq 2 | \theta_j, b_{i2}, \alpha_i) = \frac{1}{1 + e^{(\theta_j - b_{i2})}}
$$

$$
P(y_{ijk} \geq 3 | \theta_j, b_{i2}, \alpha_i) = \frac{1}{1 + e^{(\theta_j - b_{i3})}}
$$

$$
\vdots
$$

$$
P(y_{ijk} \geq k + 1 | \theta_j, b_{i2}, \alpha_i) = 0
$$

and these cumulative probabilities lead to the graded response model, or the probability of a

$$
P(y_{ijk} = k | \theta_j, b_{ik}, \alpha_i) = p(y_{ijk} \geq k | \theta_j, b_{ik}, \alpha_i) - p(y_{ijk} \geq k + 1 | \theta_j, b_{ik}, \alpha_i)
$$
\[ \frac{1}{1 + e^{(\theta_j - b_l k)}} \cdot \frac{1}{1 + e^{(\theta_j - b_l k + 1)}} \]  

(2.4)

where \( k = 1 \ldots K \), \( P(y_{ijk} \geq k | \theta_j, b_{lk}, a_i) \) is the cumulative probability of scoring in or above category \( k \) of item \( i \) given \( \theta_j \) and the item parameters \( a_i \) as before is the item slope \( b_{lk} \) is the category boundary or threshold for category \( k \) of item \( i \). The cumulative function for the middle categories looks very much like the 2PL model except for multiple \( b_{lk} \) parameters.

Thus, equation \( P(y_{ijk} \geq k | \theta_j, b_{lk}, a_i) \) is the form of the GRM model. The plots of the boundary probabilities and the probabilities of responding at a specific category in an item \( P(y_{ijk} = k | \theta_j, b_{lk}, a_i) \) are displayed in Figure 5. They are referred to as the item operating characteristic function (OCC) and the item category characteristics functions (ICC), respectively.

The OCC in top panel are the same as the two parameter logistic model for the dichotomous Items. The top curves in the (OCC) specify the probability of a response in the categories above or below the threshold. The lower panel curves (ICC) show the probability of each score categories 1, 2, 3, and 4 for a person at a specific \( \theta_j \) level. The OCCs cross the .5 probability at the point equal to the step difficulty (threshold) and their slopes are steepest at that point, although the two ICC curves (bottom) for the lowest and highest categories (1 and 5) cross the .5 probability line. The peaks of the curves do not have any obvious connection to the \( b_{lk} \) parameters. We can identify which categories are less likely to be chosen from the ICC curves.

For polytomous items, the questions about the \( b_{lk} \) parameters and the \( a_i \) parameters should be: “What is the spread of the category difficulties?” and “How discriminating is each item?” If the b-parameters of an item are spread out, the item can measure across a wider range of \( \theta \). If the locations are close together or span a narrow area, this item may not differentiate well among respondents across the area. Also low discriminating items have very flat ICC curves.

**2.2 Model Assumptions**

Key assumptions in an IRT model are (1) Unidimensionality of the latent traits and (2) Local independence and (3) Monotonicity. The first assumption implies that the items share a common primary construct and that the model creates a single \( \theta_j \) for each respondent. This means that the
items collectively measure a unique underlying latent trait for each examinee and that only one latent trait influences the item responses.

Other factors affecting these responses are treated as random error (DeMars, 2010, p. 38). This is a strong assumption and may not be reasonable in many situations as tests or survey instruments may be designed to measure multiple traits. However, as previously stated, we restrict our discussion to this unidimensionality assumption and refer to the IRT models that meet this assumption as unidimensional IRT (UIRT) models. When the assumption does not hold, estimates of parameters and standard errors may be questionable.

**Local independence** indicates that if the assumption of unidimensionality holds, and then the response of a subject to one item will be independent of his or her response to another item conditional on the latent trait. This assumption also called conditional independence is closely related to the dimensionality assumption. Local independence is typically defined such that after controlling $\theta$, no substantial relationship remain between pairs if a set of items contains many locally dependence item pairs this indicates the present of multidimensionality that is not accounted by the model (Thissen and Steinberg, 2010) which may suggest that a unidimensional model is not appropriate.

In other words, if items are locally independent, they will not be correlated after conditioning on $\theta_j$ (DeMars, 2010).

Letting $y_i = (y_{i1}, y_{i2}, ..., y_{ij}, ..., y_{ijn})$ with $i= 1, ......I$ and $j=1, .....n$ be the vector of I observed response from $j^{th}$ subject having an ability $\theta_j$ the assumption of local independence can be expressed as

$$P(y_i|\theta_j, \gamma) = P(y_{i1}|\theta_j, \gamma_1)P(y_{i2}|\theta_j, \gamma_2) ... P(y_{ijn}|\theta_j, \gamma_1)$$

$$= \prod_{i=1}^{I} P(y_{ij}|\theta_j, \gamma_1)$$

$$= \prod_{i=1}^{I} f(y_{ij}|\theta_j, \gamma_i)$$

(2.5)
Where $P(y_{ij}|\theta_j)$ is the probability that the vector of observed item score for a person with trait level $\theta_j$ has the pattern $y_j$ and $\prod$ is the symbol for the product of individual probability $P(y_{ij}|\theta_j)$ for a person with trait levels $\theta_j$ obtaining of $y_{ij}$ of item i expressing equation in terms of the $\theta_j$ we obtain the likelihood function

$$L(y_{ij}|\theta_j, \gamma_i) = \prod_{i=1}^{I_j} P(y_{ij}|\theta_j, \gamma_i)$$

(2.6)

This can be generalized to the probability $P(y|\theta)$ of a complete set of response from n person to I items an instrument where $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ represent the vector of latent score for all respondent as shown in equation

$$P(y|\theta, \gamma) = \prod_{i=1}^{n} P(y_j|\theta_j, \gamma) = \prod_{j=1}^{n} \prod_{i=1}^{I_j} f(y_{ij}|\theta_j, \gamma_i)$$

(2.7)

The Monotonicity assumption determine the probabilistic relationship, $p_r(Y_j = 1|\theta) = p_j(\theta)$, between positive response to items $j$ and latent proficiency $\theta$. $p_j(\theta)$ is also called the items response function (IRF). Thus, the examinee with the higher proficiency have a higher probability of answering item correctly. A considerable number of models has been proposed for IRT models, and they fall in two classes according to the estimation method of the IRF, $p_j(\theta)$ for binary response data the number of parameter determine the shape of the IRF and characterizes IRT model, as Rasch model (or one parameter model), the two parameter model and the three parameter model for polytomous response data (i.e. data with more than two categories) the model include graded response model, Partial credit model (including the generalized partial credit model). IRT models are commonly fitted with maximum marginal likelihood estimation (Bock and Aitkin, 1981) for which usually assumed that

- Item response function are logistic and
- The ability distribution of population follows a parametric distribution (i.e. normal or uniform).
3 Programs for the International Assessment of Adult Competencies (PIAAC).

3.1 Data Description

The Survey of Adult Skills (PIAAC) collected competency (cognitive) information through a series of assessment booklets containing literacy, numeracy and Problem solving in technology-rich environments (PSTRE) and descriptive information through a background questionnaire (BQ). PIAAC respondent ages ranged from 16 to 65. Eligible participants included individuals who were living in households, institutional residents were excluded. Twenty-four countries participated in PIAAC and the sample size varies across the countries.

<table>
<thead>
<tr>
<th>Country</th>
<th>Sample size(n)</th>
<th>Country</th>
<th>Sample size(n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>7,430</td>
<td>Italy</td>
<td>4,621</td>
</tr>
<tr>
<td>Austria</td>
<td>5,130</td>
<td>Japan</td>
<td>5,278</td>
</tr>
<tr>
<td>Canada</td>
<td>27,285</td>
<td>Korea</td>
<td>6,667</td>
</tr>
<tr>
<td>Cyprus</td>
<td>5,053</td>
<td>Netherland</td>
<td>5,170</td>
</tr>
<tr>
<td>Czech republic</td>
<td>6,102</td>
<td>Norway</td>
<td>5,128</td>
</tr>
<tr>
<td>Denmark</td>
<td>7,328</td>
<td>Poland</td>
<td>9,366</td>
</tr>
<tr>
<td>Estonia</td>
<td>7,632</td>
<td>Russian Federation</td>
<td>3,892</td>
</tr>
<tr>
<td>Finland</td>
<td>5,464</td>
<td>Slovak republic</td>
<td>5,723</td>
</tr>
<tr>
<td>Belgium</td>
<td>5,464</td>
<td>Spain</td>
<td>6,055</td>
</tr>
</tbody>
</table>

Table 1. Participating countries in PIAAC and sample sizes.

For this thesis we select only two countries: Sweden and United States. Both completed the Problem solving in technology-rich environments (PSTRE) cognitive assessment domain. The
PSTRE domain is organized around three core dimensions: the cognitive strategies and processes a person uses to solve a problem, the tasks or problem statements that trigger and condition problem solving, and the technologies through which the problem solving is conducted.

In the Survey of Adult Skills, problem solving in technology-rich environments is defined as “using digital technology, communication tools and networks to acquire and evaluate information, communicate with others and perform practical tasks”. The Survey of Adult Skills focuses on “the abilities to solve problems for personal, work and civic purposes by setting up appropriate goals and plans, and accessing and making use of information through computers and computer networks”.

The problem solving in technology-rich environments domain covers the specific types of problems people deal with when using ICT. These problems share the following characteristics:

- The problem is primarily a consequence of the availability of new technologies.
- The solution to the problem requires the use of computer-based artefacts (applications, representational formats, computational procedures).

The problems are related to technology-rich environments themselves (e.g. how to operate a computer, how to fix a settings problem, how to use an Internet browser). Problem solving in technology-rich environments is a domain of competency that represents the intersection of what are sometimes described as “computer literacy” skills (i.e. the capacity to use ICT tools and applications) and the cognitive skills required to solve problems. Some basic knowledge regarding the use of ICT input devices (e.g. use of a keyboard and mouse and screen displays), file management tools, applications (word processing, e-mail), and graphic interfaces is essential for performing assessment tasks. However, the objective is not to test the use of ICT tools and applications in isolation, but rather to assess the capacity of adults to use these tools to access, process, evaluate and analyse information effectively.

The PSTRE domain consists of 14 items, ten items are dichotomous and four items are polychotomous scored with four response categories coded as 0= incorrect, 1= partially correct 1, 2= partially correct 2, 3= correct.
<table>
<thead>
<tr>
<th>Items</th>
<th>ID</th>
<th>Description</th>
<th>Cognitive skill</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02X00s</td>
<td>Using information from a novel Internet application and several e-mail messages, establish and apply criteria to solve a scheduling problem where an impasse must be resolved, and communicate the outcome</td>
<td>Problem solving</td>
</tr>
<tr>
<td>2</td>
<td>U04A00s</td>
<td>Using information embedded in an e-mail message, establish and apply the criteria to transform the e-mail information to a spreadsheet. Monitor the progress of correctly organising information to perform computations through novel built-in functions.</td>
<td>Problem solving</td>
</tr>
<tr>
<td>3</td>
<td>U11b00s</td>
<td>Infer the proper folder destination in order to transfer a subset of incoming e-mail messages based on the subject header and the specific contents of each message.</td>
<td>Problem solving</td>
</tr>
<tr>
<td>4</td>
<td>U23X00s</td>
<td>Enact a plan to navigate through a website to complete an explicitly specified consumer transaction. Monitor the progress of submitting a request, retrieving an e-mail message, and filling out a novel online form.</td>
<td>Problem solving</td>
</tr>
</tbody>
</table>

Table 2. Four items used for the analysis

International studies in education such as PIAAC provide the information on how adults from different countries with similar and dissimilar educational environment perform on test and researches the factor influence student achievement. The following figure displays the raw score of female (group 2) and male (group 1) adult’s answer for the question problem solving domain by country.
Figure 6. Top US and bottom Sweden. Bar plots display 1 = male, 2 = female and the raw score of female and male adults answers for the question problem solving domain by country. Four categories in each item.

3.2 Assessment of Model Assumption

As with most statistical procedures, analysis based on IRT has several underlying assumptions. The primary assumptions of the 2PLM and GRM have been briefly discussed in section 2.2: they are unidimensionality, local independence and monotonicity (functional form). As these assumptions are never precisely met by real data sets the test developer can be confident the result of an IRT analysis to the degree that these assumption holds (Embreston and Reise, 2000).

**Dimensionality** In IRT modeling it is assumed that one or more variables explain the relationship among individual response items (MC Donald 1999). Most IRT models are predicted on the assumption that the observed relationship among item response can be fully explained by a single continuous (usually normally distributed) individual construct. Unfortunately no real test is
purely unidimensional. As a result there has been a lot of debate in the IRT literature concerning the wide variety of dimensionality assessment methods. Generally speaking observed data are often consistent with a number of different dimensions and the true dimensionally is not known, however some researcher argued that unidimensionalty of IRT is acceptable if the researcher first provides a strong factor in the data.

A common approach to address the dimensionality problem is to use a combination of exploratory and confirmatory factor analysis designed for ordinal data prior to fitting an IRT model. One approach that has been proposed as a frame work addressing this problem is use bifactor modeling (Reise, S.P., Morizot, J., & Hayes, R.D.,2007). In a bifactor model all items load on to general factor that is expected to represent the construct of interest. In the most interpretable types of bifactor structure each item also loads onto one and only one group factor and the general and group factors are constrained to be orthogonal. The general factor is believed to reflect the construction the researcher intended to measure whereas the groups factor represent multidimensionality in the data.

The bifactor model can be compared to a unidimensional model; if the data are unidimensional enough for IRT, the factor loading on the general should be similar to the factor loading obtained from the unidimensional model and any substantial difference shows forcing multidimensional data in to unidimensional data which results in an invalid solution. In addition we assessed dimensionality by scree plot of the eigenvalues of the polytomous items correlation matrix. The plot in Figure 7 appears to have one dominant eigen values indicating that the items may cluster in to one factor,that is, unidimensional may be appropriate.

Dimensionality Assessment Exploratory item analysis was conducted on the four items to investigate the dimensionality structure of the data, first a matrix of polychoric correlation was computed. Upon examination of the correlation matrix it was evident that all items are correlated, exploratory items factor analysis was performed on the polychoric correlation using ordinary least square extraction and oblique quartimax rotation. Oblique quartimax rotation is equivalent to direct quartimin rotation (Browen and Cudeck ,1993) and has been endorsed by several prominent factor analysts (Browen and Cudeck ,1993). For the PSTRE, as seen in Table 2. One dimensional factor solutions were explored. Factor loadings (pattern coefficients) of .40 or larger were considered to be meaningful.
Figure 7. Top US and Bottom US scree plot of the eigen values, the scree clearly begins to flatten at the second eigen value.
As seen in Table 4 the factor loadings in the unidimensional solution were not forced in to a unidimensional IRT model the factor loadings in the unidimensional solutions were similar to the general factor loading in the bifactor. These indicates that the fitting a unidimensional IRT modeling to these data would not be expected to produce substantial distortion of IRT slope parameter estimates. The result of the various dimensionality assessment methods collectively suggested the data were unidimensional enough for interpretation of the IRT model.

The local item independence diagnostic statistics (referred to as LD) was developed by Chen and Thissen (1997) proposed the LD $X^2$ statistic, computed by comparing the observed and expected frequencies in each of the two-way marginal cross tabulation between responses to each item and each of the other item. These diagnostic are (approximately) standardized $X^2$ values (they are approximately z-score) that become large if a pair of items indicates local dependence. The values greater than 10 on this LD statistic indicate substantial local item dependence (Cai,Le and Thissen,2011). The IRT analysis was conducted on the four items using GRM and as seen in Table 3 the local independence criterion (LD values less than 10) was met for all items pairs.

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$X^2$</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>0.0</td>
<td>3.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>0.0</td>
<td>1.1</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>0.5</td>
<td>7.4</td>
<td>8.2</td>
<td>9.4</td>
</tr>
<tr>
<td>Item</td>
<td>Label</td>
<td>$X^2$</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>0.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>0.1</td>
<td>4.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>0.0</td>
<td>4.5</td>
<td>4.0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>0.4</td>
<td>2.3</td>
<td>-0.6</td>
<td>1.9</td>
</tr>
</tbody>
</table>

Table 3. Top US and bottom Sweden Standardized LD $X^2$ Statistics

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$\lambda_1$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>0.85</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>0.76</td>
<td>0.10</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>0.70</td>
<td>0.06</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>0.73</td>
<td>0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$\lambda_1$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>0.78</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>0.59</td>
<td>0.09</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>0.61</td>
<td>0.06</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>0.78</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table 4. Top US and Sweden Factor Loadings (Unidimensional and bifactor model)
4 Data Analysis and Result

4.1 Goodness of fit tests

The GRM was fitted to the four items for two countries separately. The M2 goodness of fit statistic (Maydeu-Olivares and Joe, 2005) is a chi square distribution statistics and was used to assess the overall IRT model data fit. The M2 goodness of fit was computed using formula 11 in appendix B. Small values of M2 indicate better model fit, but some lack of fit is generally expected when strong parametric IRT models such as the GRM are applied to real data (Cai et al. 2011). For this reason, the M2 statistic was supplemented by the root mean square error of approximation (RMSEA) statistic, this index is not as sensitive to sample size or over parameterization as chi square distributed statistics (Browen and Cudeck, 1993) and is routinely used to assess the fit of structural equation models.

Maydeu-Olivaries (2005) and Cai, (2011) demonstrate that RMSEA tends to yield similar interpretation whether a confirmatory factor analysis or IRT model is applied. Specific cutoff values for interpreting RMSEA are not fully agreed upon however, we chose to use a cut off 0.06 based on the work of Hu and Bentler (1999); thus, values less than 0.06 indicate acceptable fit and larger values indicates misfit of IRT model. Appendix Table A. 1. and Table A. 3 display that the RMSEA 0.03 and 0.04 for US and Sweden which are less than the cut off value 0.6 and indicate that the data fits the model well.

Trace line fits were also assessed the functional form assumption we evaluated using the $s - X^2$ statistic calculated by using formula 10 in Appendix B. (Orlando and Thissen, 2000; 2003) and a nominal alpha level of 0.05 for each items. The $s - X^2$ statistic summarizes the relationship between the trace line (S) predicted by the model and the empirical lines (s) based on summed scores a statistically significant value indicates item model misfit because the item response function predicted by the model differs from observed in the data. If the p values less than .05 we would have evidence against the null hypothesis of the fitted model. Otherwise, we would fail to reject $H_0$ and conclude that we do not have enough evidence to reject the possibility of $H_0$. 
<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$X^2$</th>
<th>d.f.</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>33.20</td>
<td>19</td>
<td>0.0527</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>19.48</td>
<td>16</td>
<td>0.2441</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>26.18</td>
<td>19</td>
<td>0.1248</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>60.89</td>
<td>15</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$X^2$</th>
<th>d.f.</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>32.44</td>
<td>23</td>
<td>0.0911</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>27.49</td>
<td>22</td>
<td>0.1927</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>28.43</td>
<td>23</td>
<td>0.1996</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>35.36</td>
<td>19</td>
<td>0.0126</td>
</tr>
</tbody>
</table>

Table 5. Top US and bottom Sweden $s - X^2$ Item Level Diagnostic Statistics.

The $X^2$ statistic tests the null hypothesis that an item fits the polytomous IRT model. A non-significant chi-square value means that the null hypothesis cannot be rejected. The summary table indicates that for all items except item 4 in both US and Sweden the trace line have been fitted sufficiently well that the model expected proportion responding 0,1,2,3 match the observed data well.

### 4.2 Item Parameter Estimation

With the likelihood conditional on $\theta = \theta_1, \theta_2, \ldots, \theta_n$ and on assumed normal distribution form $g(\theta_j|\psi)$ for the independent and identically distributed latent trait the marginal log likelihood is

$$ l(\gamma) = \sum_{i=1}^{l} \sum_{j=1}^{n} log f(y_{ij}|\theta_j, \gamma_i)g(\theta_j|\psi) \ d\theta_j $$

(4.1)
where \( \psi \) is a set of hyper parameter for mean and standard deviation, usually set at 0 and 1 respectively, because of the assumption of local independence mentioned previously, maximization of formula 4.1 Reduces to maximization of

\[
I_i(y_i) = \sum_{j=1}^{n} \log f(y_{ij} | \theta_j, y_i)g(\theta_j | \psi) \, d\theta_j
\]  

(4.2)

For one item at a time where \( \gamma_i = (a_i, b_i) \), \( i=1, \ldots, I \), and the integrals are numerically approximated using a Gauss-Hermite quadrature algorithm. After the item parameters are estimated, they are used to update information on the distribution of \( \theta \) and the item parameters are re-estimated. The procedure is repeated until the estimated values stabilize or converges.

### 4.3 Test characteristic curve

The test characteristic curve graphically illustrates the relationship between \( \theta \) and predicted summed scores on a test predicted summed scores are analogous to CTT (Classical test theory) estimated true and are computed using the IRT properties of the items on the test (Hambleton). This computation is possible because IRT model probability of selecting each response option for an item additionally, each option is associated with specific values (the same types of whole number values that is typically used to compute observed summed scores).

For a dichotomous item, the non-keyed and keyed options (i.e. what the examinee has keyed in) are assigned values of 0 and 1 respectively as previously discussed when polytomous items are analyzed with IRT the options are assigned values 0...k-1. The test characteristics curve is a linear combination of the trace lines for all the items (Yen and Fitzpatrick, 2006).
Figure 8. Top US and bottom Sweden Test characteristic curves,
one curve for each item category, 0,1,2,3.
<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$a$</th>
<th>s.e.</th>
<th>$b_1$</th>
<th>s.e.</th>
<th>$b_2$</th>
<th>s.e.</th>
<th>$b_3$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>2.77</td>
<td>0.41</td>
<td>0.67</td>
<td>0.06</td>
<td>0.99</td>
<td>0.07</td>
<td>1.45</td>
<td>0.09</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>2.00</td>
<td>0.36</td>
<td>1.31</td>
<td>0.12</td>
<td>1.35</td>
<td>0.12</td>
<td>1.54</td>
<td>0.14</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>1.68</td>
<td>0.17</td>
<td>0.66</td>
<td>0.07</td>
<td>0.72</td>
<td>0.07</td>
<td>1.07</td>
<td>0.09</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>1.81</td>
<td>0.19</td>
<td>-0.14</td>
<td>0.06</td>
<td>-0.12</td>
<td>0.06</td>
<td>0.30</td>
<td>0.06</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$a$</th>
<th>s.e.</th>
<th>$b_1$</th>
<th>s.e.</th>
<th>$b_2$</th>
<th>s.e.</th>
<th>$b_3$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>2.08</td>
<td>0.22</td>
<td>0.16</td>
<td>0.04</td>
<td>0.58</td>
<td>0.05</td>
<td>1.21</td>
<td>0.08</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>1.24</td>
<td>0.17</td>
<td>1.22</td>
<td>0.12</td>
<td>1.31</td>
<td>0.13</td>
<td>1.65</td>
<td>0.16</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>1.32</td>
<td>0.11</td>
<td>0.34</td>
<td>0.06</td>
<td>0.43</td>
<td>0.06</td>
<td>1.08</td>
<td>0.08</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>2.14</td>
<td>0.24</td>
<td>-0.15</td>
<td>0.04</td>
<td>-0.12</td>
<td>0.04</td>
<td>0.34</td>
<td>0.05</td>
</tr>
</tbody>
</table>


The trace lines for four items of the US and Sweden are shown in Figure 8. The trace line can be viewed as the regression of item score on the underlying variable θ. Each curve depicts each response category. The Figure 8 brings out which categories are less likely to be chosen. The more response categories, more ambiguous the figure will be.

Taking a look at US item 1 in Figure 8, the probability of each of four response categories were presented in the left-most panel. The curve of the score 0 is a decreasing function of θ and it crosses the 0.5 probability line at $b_{l_1} = 0.6$. The curve of the highest score category $k=3$ crosses the 0.5 probability line at $b_{l_3}=1.45$. The value of the other items has the same characteristics. Items 1 and 2 seem the two hardest item with highest $b_{l_3}=1.45$ and 1.54 respectively. With increasing discrimination parameters $\alpha$ the curve will be steeper for the extreme score categories ($k = 0$ or $m_i$). Meanwhile, the curves for the categories between the extremes ($k = 1..m_i-1$) become more peaked. The values $\alpha$ of the first and last item for US are larger than the others.
Hence, we could see the probabilities curves item 1 is more peaked than those of the others. The dashed line in each panel in Figure 8 is the information curve of each item; more about the information curve will be discussed in the next part.

Generally, as parameter $\alpha$ increases, the probability of getting a specific score changes more quickly with a change in $\theta$. The values $b$ of the highest score categories of items 3 and 4 are low. Items 3, 4 seem to be easy, items 1 and 2 have the highest $b_{i3} = 1.45$ and 1.54 which mean they are more difficult than the other ones. The model does restrict the items to be at an ordinal measure level. An item is ordinal if the estimated log-linear effects ($b$) of the trait on the observed variable all increase (or all decreases) numbers the table 6 therefore shows ordinality holds for all observed variables.

The same effects are also unevenly spaced, in some cases such as two and three (partially correct 1 and partially correct 2) for the item U04a000S and U23x000S in Sweden the difference is not significant for two different categories. This suggests that these categories represent much the same and that there for these items cannot be taken to be interval measure level.

Table 6 shows also that the parameter has been estimated with considerable uncertainty which includes sampling design and interviewer effects, since we have included these in the model estimation procedure .In spite of the larger standard error most coefficients have been estimated with sufficient precision to distinguish between the parameter values. The highest uncertainty is related with the highest categories and the highest item difficulty.

### 4.4 Test information Curve

The test information curve represents the amounts of information that a test including all items, provides for estimating each individuals location on the theta scale (de Ayala, 2009). The shape of this function is directly related to the IRT parameter of a set of test items test information is greatest where the test items are concentrated (i.e. more values are located ) and the items are more related to theta (i.e. the a values are larger) a test differentiates between individual of different theta levels at the points which test information is the highest a test differentiates less well at theta levels where test information is lower it is represented by the dotted lines in graph a more direct measures of the quality than has been used so far is the items information it is the inverse of the error variance of the maximum likelihood estimate of the traits that one can get from each item ,and can be seen as a generalized reliability.
The term information in IRT describes how certain we feel about the estimate of person’s location $\theta_i$ and is based on the fisher information matrix.

$$I(\theta) = E\left(\frac{\partial \ln L}{\partial \theta}\right)^2 = E\left(\frac{\partial^2 \ln L}{\partial \theta^2}\right)$$  \hspace{1cm} (4.3)

where $L$ is the likelihood function defined in 2.4 individually each dichotomous function is defined as

$$I_i(\theta) = -E\left(\frac{\partial^2 \ln p_i(\theta)}{\partial \theta^2}\right)$$  \hspace{1cm} (4.4)

where $p_i(\theta)$ is the probability of a correct response to dichotomous item $i$. For polytomous items, the information function $I_i(\theta)$ is the sum of the category information function $I_{ik}(\theta)$ which is based on the Fisher information matrix and defined by Samejima(1969) as in equation

$$I_{ik}(\theta) = \frac{\partial^2 \ln p_{ik}(\theta)}{\partial \theta^2} = -\frac{\partial}{\partial \theta}\left(\frac{p'_{ik}(\theta)}{p_{ik}(\theta)}\right) =$$

$$\frac{(p'_{ik}(\theta))^2 - p_{ik}(\theta)p''_{ik}(\theta)}{(p_{ik}(\theta))^2}$$  \hspace{1cm} (4.5)

where $p_{ik}(\theta)$ is the probability of response in category $k$ to item $i$ defined in equation and $p'_{ik}(\theta)$ and $p''_{ik}(\theta)$ are the first and second derivatives of $p_{ik}(\theta)$. Thus the information function is calculated as

$$I_i(\theta) = \sum_{k=1}^{K} I_{ik}(\theta)p_{ik}(\theta)$$  \hspace{1cm} (4.6)

The total information is sum of the item information

$$I(\theta) = \sum_{i=1}^{I} I_i(\theta)$$  \hspace{1cm} (4.7)

In the Figure 9 the information has been plotted on a log scale to allow for comparison of the different items which vary widely in information provided.
Figure 9, The top US and bottom Sweden, Test information Curves

The information function (Figure 9); this function indicates the precision of measurement for persons at different levels of proficiency. The shape of IF is dependent on the item the higher items discrimination and the more peaked the IF will be.

Figure 9 indicates information function curve (IFC) for each item; here we can see the relationship between the information of various item and proficiency levels. The IFCs of item 2 and 3 are almost similar. Interestingly, item 1 provides more information for examinees at the highest proficiency level the curve spreads in the most area of high proficiency. Item 2 and 3 provide least information whereas item 4 is distributed uniformly across the proficiency range.
4.5 Test standard error curve

The test standard error curve represents the amounts of measurement error a test contains controlling for $\theta$. It is computed as the inverse of the information curve. If the Information curve is low, there are large errors in estimating an individual’s location on the theta scale. Conversely when the curve is higher there is better error and thus less precision in estimating individual theta scores. The theta at which the standard error function is the lowest corresponds to the peak of the test information function. Thus, IRT measure precision for a test is a maximized for the theta (or range of theta) in which information is the highest and error is the lowest.

Figure 10. Top US bottom Sweden standard error curves
For example, item 1 yields its maximum information for estimating $\theta$ at its difficulty of 1.45 (Table 6). Hence, using this item to estimate persons with big $\theta$ would not provide precise estimates and the SE would be large. Total information function (TIF) is calculated as the sum of all item information functions from the scale with the formula 4.7. The variance of the ability estimate can be estimated as the reciprocal value of TIF. The SE is equal to the square root of the variance Figure 10 shows the TIF as well as the SE for four items. The TIF is displayed in line and its values can be read from the left-hand axis. The SE is displayed in dotted line, and its values can be read from the right-hand axis. The TIF indicates how well the entire instrument can estimate the proficiency $\theta$. The maximum information gained in US and Sweden around 1 and 0.5 proficiency levels respectively. The instruments provide less information for large $\theta$. At little skewed ability values we achieved the most information and the smallest standard error for both US and Sweden.

4.6 Reliability Analysis

Reliability is the correlation between the observed variable and the true score when the variable is an exact or imprecise indicator of the true score (Cohen 1988). Inexact measure may come from guessing, differential perception, recording errors, etc. on the part of the observers. Cronbach’s $\alpha$ is a coefficient of reliability or internal consistency of latent construct. It measure how well a set of items or variables measures a single unidimensional latent construct. When data have a multidimensional structure Cronbach’s $\alpha$, is calculated with formula

$$\alpha = \frac{m}{m-1} \left( 1 - \frac{\sum_{j=1}^{m} s_j^2}{s_y^2} \right)$$

Where, $s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2$

$$\bar{Y} = \sum_{j=1}^{m} \bar{X}_j$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \bar{Y})^2$$

$n$ is the number of observation

$m$ is the number of items

$s_j^2$ is the variance of items $j$
$s_y^2$ is the variance of total score

The standardized Cronbach’s $\alpha$ can be written as a function of the number of items $m$ and the average inter correlation $\bar{r}$ among the items.

$$\alpha = \frac{m \bar{r}}{1 + (m-1) \bar{r}}$$

From this formula one can see that if one increases the number of items Cronbach’s $\alpha$ will raise. Additionally if the average inter-item correlation is low alpha, will be low. The range of the alpha is from 0 to 1. A reliability coefficient of 0.70 or higher is considered acceptable in most research situations. Reliability measures the precision of test and can be characterized as a function of proficiency $\theta$. The marginal reliability is

$$\bar{\rho} = \frac{\sigma_\theta - \sigma_\epsilon^2}{\sigma_\theta^2}$$

(4.9)

$$\bar{\sigma}_\epsilon^2 = \int \bar{\sigma}_\epsilon^2 g(\theta) d\theta$$

where $\bar{\sigma}_\epsilon^2$ is the marginal error variance, $g(\theta)$ is proficiency density is fixed as N(0,1), $\sigma_\theta^2$ is the expected value of the error variance function can be calculated from the information $I(\theta)$. In this section, the coefficients for reliability analysis were computed for each country with the formulas 4.8 and 4.9:

- Cronbach’s $\alpha$ coefficient based on covariance
- The average inter item correlation $r$

Table 7 shows the Cronbach’s $\alpha$ coefficient in all country are greater than the cut point 0.7 for US and Sweden. However the result is not far from 0.7, reliability is high which means a set of items measure a single unidimensional latent construct (ability of examines) very well.
<table>
<thead>
<tr>
<th>Item</th>
<th>Response Average</th>
<th>Std. Dev.</th>
<th>Item-Total Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.590</td>
<td>1.038</td>
<td>0.5683</td>
</tr>
<tr>
<td>2</td>
<td>0.425</td>
<td>1.011</td>
<td>0.4466</td>
</tr>
<tr>
<td>3</td>
<td>0.846</td>
<td>1.289</td>
<td>0.5021</td>
</tr>
<tr>
<td>4</td>
<td>1.445</td>
<td>1.425</td>
<td>0.4798</td>
</tr>
</tbody>
</table>

With Item Deleted

<table>
<thead>
<tr>
<th>Item</th>
<th>Response Average</th>
<th>Std. Dev.</th>
<th>Item-Total Correlation α</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.075</td>
<td>1.208</td>
<td>0.6398</td>
</tr>
<tr>
<td>2</td>
<td>0.645</td>
<td>1.181</td>
<td>0.4483</td>
</tr>
<tr>
<td>3</td>
<td>1.184</td>
<td>1.358</td>
<td>0.5492</td>
</tr>
<tr>
<td>4</td>
<td>1.483</td>
<td>1.394</td>
<td>0.6334</td>
</tr>
</tbody>
</table>

Table 7. Top US bottom Sweden Reliability analysis of the problem solving test

Coefficient alpha: 0.7041 and 0.7824 for US and Sweden respectively.
4.7 Differential Item Functioning (DIF)

The graded response model is applied to real data from the program for international adult assessment (PIAAC) to investigate the difference of groups and item analysis. In this study, gender differences are explained by the difference in latent trait probability across gender using the graded response model. One of the advantages of this approach is that the group comparison and item analysis can be done simultaneously another advantage is that model can separate group differences from item effects that otherwise might be confounded.

Differential item function (DIF) detecting can be confounded when the distributions of the groups are different. For example, suppose that the distribution of men’s math ability has a bigger variance and a higher mean than the distribution of women’s ability. Then, items can be determined to be DIF items due to group difference even if they are good items the error associated with an items as DIF is shown as type I error. This section presents an analysis of PIAAC data address the following research questions.

1. Is incorporating, covariates (gender) in to GRM model useful to identify group difference Interims of the measured proficiency?

2. Do items perform differently across groups after taking in to account the difference Interims of the measured proficiency?

4.7.1 Analysis for comparison of countries

International studies in education provide the information on how adults from different countries with similar and dissimilar educational environment perform on test. Figure 11 displays the average for problem solving for US and Sweden.
Averages for problem solving, adults (16-65) by Person resolved gender from EQ and QC check (derived) [GENDERR] for jurisdiction and year: PIAAC 2012 2012

![Bar chart showing average problem solving scores for US and Sweden.]

NOTE: The Problem solving scale ranges from 0 to 500. Some apparent differences between estimates may not be statistically significant. SOURCE: Organization for Economic Cooperation and Development (OECD), Program for the International Assessment of Adult Competence (PIAAC). 2012.

Figure 11. Average for problem solving for US and Sweden

Cross country comparison based on PIAAC are typically reported in terms of means and variance of the scale scores by country and the rank are reported as means score of national achievements. However, it is worth investigating why the average achievements differ. In problem solving, US Adults is lower than Adults in Sweden. Sweden ranked above the US and reason could be the absence in the US of the best performing or an abundance of poorly performing adults. The IRT model considers the whole range of distribution of ability across the grouping variables (e.g., gender).

The proposed model is applied to investigate how the problem solving skill performance of US is lower than Sweden and to examine whether there is a gender different in terms of problem solving ability measured by four polytomous items. The analysis is performed by country to see whether there is a gender effect with in that country.

The result is attached in the appendix table with for each country based on the AIC, BIC and the RMSEA and compare these two models in terms of model fit. It starts with US and then goes through Sweden.
US, as display in appendix Table A. 1 and A.2 we fit GRM without gender which result in BIC = 6037.54, AIC = 5937.27 and RMSEA 0.03. Then the GRM fitted to the data using the gender covariate BIC = 6177.12, AIC = 5964.06 and RMSEA 0.02.

Sweden, as display in appendix Table A. 2 and A.4 GRM gender BIC= 9881.62, AIC = 9779.14 and RMSEA = 0.04. Then the GRM fitted to the data using the gender covariate BIC = 10016.43 AIC = 9798.66 And RMSEA 0.02.

In general, the GRM model including gender as covariate fits the data well with small RMSEA and it indicates that there is gender difference in problem solving ability in both countries US and Sweden.

<table>
<thead>
<tr>
<th>Item numbers in:</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Total $X^2$</th>
<th>d.f.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2.0</td>
<td>4</td>
<td>0.7294</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0.5</td>
<td>4</td>
<td>0.9712</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1.0</td>
<td>4</td>
<td>0.9068</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2.4</td>
<td>4</td>
<td>0.6590</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item numbers in:</th>
<th>Group 1</th>
<th>Group 2</th>
<th>Total $X^2$</th>
<th>d.f.</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2.6</td>
<td>4</td>
<td>0.6250</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2.6</td>
<td>4</td>
<td>0.6265</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4.3</td>
<td>4</td>
<td>0.3695</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>3.6</td>
<td>4</td>
<td>0.4682</td>
</tr>
</tbody>
</table>
Table 8. Top US and bottom Sweden, DIF Statistics for Graded Items

Under the hypothesis of perfect fit model fit the $s - \chi^2$ statistic is approximately distributed as $\chi^2$ value with tabulated degree of freedom significance value indicate a lack of fit, the statistics tabulated for GRM model including gender as covariate fit illustrate all item are non-significant at the P=0.05 level.

![Graphs of information curves for US and Sweden](image)

Figure 12. Sweden top group 1 and bottom group 2 total information curves

Look as example at Figure 12; it displays that there is the difference in shape of the information curve which indicates that items perform differently across the groups. The same difference is also seen in the US (Appendix A Figure4). In addition, Tables A: 5, 6, 7 and 8 in Appendix A displays the distribution of item discrimination value ($\alpha$) is the same with IRT model without gender difference discrimination value ($\alpha$) which means the highest discriminate item is item 1 and the last is item 3 the highest b value is associated with item 1 and 2. However, items perform differently across the groups males and females. For example, let us look at item 1 (Tables A: 5 and 6) which has a higher $\alpha$ values for group 2 than 1 in US the reverse is true in Sweden. In conclusion, there is a gender difference in both Sweden and US. In general; the items are more difficult for females than males in the US, whereas they are less difficult for females than males in Sweden.
5 Discussion

The goal of this study was to show how measurement models can be used in an educational comparative survey, and in particular to demonstrate the use of a latent trait model for analysis of the quality of an item. We have formulated an IRT model applied to the PIAAC data, furthermore we compare quality of item for two countries, which has been previously estimated US adult population has a low score in problem solving ability than has Swedish adult population. The study examined a comparison of the proficiency of the two countries interims of problem solving on the four polytomous items showing that the proficiency distribution is different over latent trait.

In this section, we will provide an overview of the obtained result and drawn several conclusions first of all, the reliability analysis determined by Cronbach’s α indicated the problem solving item measured the single latent construct well. The value of the Cronbach’s α has not increased when almost any of the items in both data sets dropped. Explanatory factor analysis was performed for two countries. The interrelationship with items have been explored and explained by one latent variable. IRT is a use full tool for both test theory and test development. Nevertheless, the assumption of local independence with items makes the model become limited in several types of test.

The polytomous item response theory was applied to data using IRTPRO 2.1. The items are locally independent. In graded response model each item is scored polytomously. The obtained outcome showed that items 1 and 2 for both countries are the two hardest items. Item 3 and 4 seems to be easy. Items 2 and 3 gives least information but on the other hand item 1 yields peak information for examinees and Item 4 is distributed uniformly across the proficiency range. Items 1, 2 as well as 3 yields much more information in positive proficiency level axis.

Given an acceptable model-fit, the method and model of IRT can provide researchers with powerful tools for revising personality test. The advantages are due to primarily IRT conceptualization of measurement precision as with IRT items level function (e.g. trace lines). IRT measurement precision for a test is typically represented graphically as a function of the latent trait and allowed to vary across the continuum of this variable (as opposed to CTT, where
all summary statistics are commonly reported). The following test level functions are related to
IRT measurement precision: the test information curve, test standard error curve, and test
characteristic curve. We investigate the quality of the question using the item characteristic
curves and information function it provides. The quality in both countries was almost similar.
The first and last item for Sweden gives the most information for test. The middle two items
gives low information for both countries. US get the most information only in the first item. The
first item has very high peak and provides much more in formation in all selected country’s about
their problem solving abilities of adults. The first and second items are not good items to provide
information about adult’s problem solving ability comparing with the other two items. The last
item has good quality overall, and covers the whole range quite well.

This finding is important. Items with an approximately equal amount of information across
the range of the trait are desirable, especially in cross national research. An item with much skew
in its information function is less likely to be a good representative items. In addition; the trace
line indicates the second response option in blue line in Figure 8 was generally not well utilized
by respondents so we combined the middle two response categories to maintain a consistent
response metric across all of the items. Thus we recommend that future researcher adopt a three
point response scale for problem solving unit (0=not correct, 1=partially correct, 2=correct) to be
useful for cross national comparisons. This is so because even if the information functions were
the same in all countries, countries with higher average would have higher measurement level
quality; as measurement errors affect the analysis of means and regression (Fuller, Wayne A.
(1987)). Differential measurement errors a cross countries invalidate comparison means and
relationships. This study also demonstrated the use of IRT model for differential item functioning
(DIF) by allowing for group effects on item response functions (IRFs) and results indicate that
item parameter estimate and IFs are not homogeneous across two groups.

Several limitation of this study may attribute to the generalizability of results. The study was
based on selected polytomous score items. The burden of the respondent could be improved by
removing some items, but specific decisions regarding which items to remove admittedly post-
hoc. Additionally, the problem solving items have been screened for differential item functioning
only for gender it is therefore unclear whether the result analysis apply equality of different
educational back ground or ethnic groups future research may explore on this issue.
References

Techniques (2nd ed.). New York: Marcel Dekker.
criteria versus new alternatives, Structural Equation Modeling, 6(1), 1-55.
Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee’s
ability.
Bollen & J. S. Long (Eds.), testing structural equation models (pp. 136-162). Newbury
information goodness-of-fit.
items and distribution of data on traditional criteria for assessing IRT’s
unidimensionality assumption. Quality of Life Research, 18, 447-460.
City: Oxford University Press.
De Ayala, R. J. (2009). The theory and practice of item response theory, the

Fuller, Wayne A. (1987), Measurement error models, John Wiley & Sons,

Glas, C. (2001). Differential item functioning depending on general covariates. In Boomsma,
A., van Duijn, M., and Snijders, T., editors, Essays on Item Response Theory, chapter 7,


IRT modeling [Computer software and manual]. Chicago, IL: Scientific Software
International.

Maydeu-Olivares and Joe (2005). Limited and full information estimation and testing in

Association.

item response theory model. Journal of Educational and Behavioral Statistics,
24:50–64.

frame work for classification and evaluation. Applied Psychological Measurement,
19:23–37.


Psychometric Monograph Supplement, No.:17. Springe.


Thissen and Steinberg, 2010 Likelihood-based item-fit indices for dichotomous
item response theory model.

Testing of item response theory models for sparse $2^p$ tables. British Journal of


measurement (pp,111-153)
## Appendix A

### Table and Figures

#### Table A. 1. GRM without covariate Likelihood-based Values And Goodness of Fit Statistics US

<table>
<thead>
<tr>
<th>Statistics based on the loglikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2loglikelihood:</td>
</tr>
<tr>
<td>Akaike Information Criterion (AIC):</td>
</tr>
<tr>
<td>Bayesian Information Criterion (BIC):</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics based on one- and two-way marginal tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_2 )</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>203.18</td>
</tr>
</tbody>
</table>

#### Table A. 2. GRM with covariate Likelihood-based Values and Goodness of Fit Statistics US

<table>
<thead>
<tr>
<th>Statistics based on the loglikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2loglikelihood:</td>
</tr>
<tr>
<td>Akaike Information Criterion (AIC):</td>
</tr>
<tr>
<td>Bayesian Information Criterion (BIC):</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics based on one- and two-way marginal tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>( M_2 )</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>180.87</td>
</tr>
</tbody>
</table>
### Table A. 3. GRM without covariate Likelihood-based Values and Goodness of Fit Statistics Sweden

<table>
<thead>
<tr>
<th>Statistics based on the loglikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2loglikelihood: 9747.14</td>
</tr>
<tr>
<td>Akaike Information Criterion (AIC): 9779.14</td>
</tr>
<tr>
<td>Bayesian Information Criterion (BIC): 9881.62</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics based on one- and two-way marginal tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_2$</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>431.13</td>
</tr>
</tbody>
</table>

### Table A. 4. GRM with covariate Likelihood-based values and Goodness of Fit Statistics Sweden

<table>
<thead>
<tr>
<th>Statistics based on the loglikelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>-2loglikelihood: 9730.66</td>
</tr>
<tr>
<td>Akaike Information Criterion (AIC): 9798.66</td>
</tr>
<tr>
<td>Bayesian Information Criterion (BIC): 10016.43</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Statistics based on one- and two-way marginal tables</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_2$</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>368.90</td>
</tr>
</tbody>
</table>
Table A. 5. Graded Model Item Parameter Estimates for Group 1, US

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$a$</th>
<th>s.e.</th>
<th>$b_1$</th>
<th>s.e.</th>
<th>$b_2$</th>
<th>s.e.</th>
<th>$b_3$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>2.58</td>
<td>0.89</td>
<td>0.83</td>
<td>0.12</td>
<td>1.15</td>
<td>0.16</td>
<td>1.55</td>
<td>0.21</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>2.13</td>
<td>0.82</td>
<td>1.35</td>
<td>0.21</td>
<td>1.38</td>
<td>0.21</td>
<td>1.59</td>
<td>0.25</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>1.65</td>
<td>0.31</td>
<td>0.73</td>
<td>0.12</td>
<td>0.82</td>
<td>0.13</td>
<td>1.15</td>
<td>0.16</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>1.83</td>
<td>0.33</td>
<td>-0.11</td>
<td>0.09</td>
<td>-0.10</td>
<td>0.09</td>
<td>0.29</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table A. 6. Graded Model Item Parameter Estimates for Group 2, US

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$a$</th>
<th>s.e.</th>
<th>$b_1$</th>
<th>s.e.</th>
<th>$b_2$</th>
<th>s.e.</th>
<th>$b_3$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>2.77</td>
<td>0.62</td>
<td>0.71</td>
<td>0.07</td>
<td>1.05</td>
<td>0.09</td>
<td>1.55</td>
<td>0.12</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>1.86</td>
<td>0.50</td>
<td>1.43</td>
<td>0.17</td>
<td>1.49</td>
<td>0.17</td>
<td>1.68</td>
<td>0.20</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>1.63</td>
<td>0.21</td>
<td>0.74</td>
<td>0.08</td>
<td>0.79</td>
<td>0.09</td>
<td>1.17</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>1.75</td>
<td>0.23</td>
<td>-0.06</td>
<td>0.07</td>
<td>-0.03</td>
<td>0.07</td>
<td>0.41</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Table A. 7. Graded Model Item Parameter Estimates for Group 1, Sweden.

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$a$</th>
<th>s.e.</th>
<th>$b_1$</th>
<th>s.e.</th>
<th>$b_2$</th>
<th>s.e.</th>
<th>$b_3$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>1.84</td>
<td>0.23</td>
<td>0.13</td>
<td>0.06</td>
<td>0.58</td>
<td>0.07</td>
<td>1.20</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>1.32</td>
<td>0.26</td>
<td>1.05</td>
<td>0.14</td>
<td>1.13</td>
<td>0.15</td>
<td>1.48</td>
<td>0.19</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>1.19</td>
<td>0.15</td>
<td>0.33</td>
<td>0.08</td>
<td>0.39</td>
<td>0.09</td>
<td>1.12</td>
<td>0.13</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>2.35</td>
<td>0.36</td>
<td>-0.17</td>
<td>0.06</td>
<td>-0.15</td>
<td>0.06</td>
<td>0.31</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Table A.8. Graded Model Item Parameter Estimates for Group 2, Sweden

<table>
<thead>
<tr>
<th>Item</th>
<th>Label</th>
<th>$a$</th>
<th>s.e.</th>
<th>$b_1$</th>
<th>s.e.</th>
<th>$b_2$</th>
<th>s.e.</th>
<th>$b_3$</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>U02x000S</td>
<td>2.36</td>
<td>0.37</td>
<td>0.08</td>
<td>0.06</td>
<td>0.49</td>
<td>0.07</td>
<td>1.13</td>
<td>0.11</td>
</tr>
<tr>
<td>2</td>
<td>U04a000S</td>
<td>1.13</td>
<td>0.24</td>
<td>1.36</td>
<td>0.22</td>
<td>1.45</td>
<td>0.23</td>
<td>1.79</td>
<td>0.28</td>
</tr>
<tr>
<td>3</td>
<td>U11b000S</td>
<td>1.43</td>
<td>0.17</td>
<td>0.25</td>
<td>0.08</td>
<td>0.37</td>
<td>0.08</td>
<td>0.96</td>
<td>0.11</td>
</tr>
<tr>
<td>4</td>
<td>U23x000S</td>
<td>1.91</td>
<td>0.25</td>
<td>-0.25</td>
<td>0.07</td>
<td>-0.20</td>
<td>0.07</td>
<td>0.27</td>
<td>0.07</td>
</tr>
</tbody>
</table>

**IRTPRO 2.1 for Windows**

IRTPRO (Item Response Theory for Patient-Reported Outcomes) is a statistical software for item calibration and test scoring using IRT. IRTPRO 2.1 for Windows is developed by Li Cai, David Thissen & Stephen du Toit. This product has replaced the four programs Bilog-MG, Multilog, Parscale and Testfact. The program has been tested on the Microsoft Windows platform with Windows7, Vista and XP operating systems.

Various IRT models can be implemented in IRTPRO, for example:
1. Two-parameter logistic model (2PL)
2. Three-parameter logistic model (3PL)
3. Graded model
4. Generalized Partial Credit model
5. Nominal model

IRTPRO implements the method of maximum likelihood for item parameter estimation, or it computes maximum a posteriori (MAP) estimates if prior distributions are specified for the item parameters IRT scores in IRTPRO can be computed using any of the following methods:

- Maximum a posteriori (MAP) for response patterns
- Expected a posteriori (EAP) for response patterns
- Expected a posteriori (EAP) for summed scores

IRTPRO supports both model-based and data-based graphical displays. Figure 4.1 shows two examples of graphical displays. More information about software IRTPRO 2.1 under http://www.ssicentral.com
Figure A. 1. Graphical examples of software IRTPRO 2.1

Figure A. 2. US top group 1 and bottom group 2 test characteristic curves
Figure A. 3. Sweden top group 1 and bottom group 2 test characteristic curves

Figure A. 4. US top group 1 and bottom group 2 total information curves
Appendix B

Applied statistical methods

1 latent variable Models

Latent variable models are prominent in social sciences such as education, sociology and psychology. A latent variable model is any models that assume the existence of latent variables that describe the interdependency among observed variables, some example of latent variables models include factor analysis, structural equation models, latent class and items response theory (IRT) models.

Bartholomew (1983) classified latent variable models into four categories based on the scale types of the latent and observed variables; they are factor analysis, latent trait analysis, latent profile analysis and latent class analysis. In latent class analysis and latent trait analysis models, observed variables (also called indicators) are dichotomous, ordinal, or nominal categories variables and their conditional distribution is assumed to be binomial or multinomial, as shown in table B.1. Latent class model differ from latent trait model where continuous underlying latent variable is replaced by discrete variables with class that define the homogenous groups of individual. For this master thesis focuses on categorical observed outcome variables response data to items or questions in a test or survey, latent trait model (also called IRT model) is manly discussed.

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Observed variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metrical</td>
<td>Categorical</td>
</tr>
<tr>
<td>Latent variable</td>
<td>Metrical</td>
</tr>
<tr>
<td>Metrical</td>
<td>Factor analysis</td>
</tr>
<tr>
<td>Categorical</td>
<td>Latent profile analysis</td>
</tr>
<tr>
<td>Latent profile analysis</td>
<td>Latent class analysis</td>
</tr>
</tbody>
</table>

Table B.1, Latent variable models

Educational measurement has mainly utilized IRT models. IRT models assume that the relationship between the probability of a positive response and the underlying latent is an increase one this functional relationship is called the item response function (IRF) (also called item characteristic curve).
2 Exploratory Factor Analysis

Exploratory factor analysis (EFA) is used to determine continuous latent variables which can explain the correlations a set of observed variables are referred to as factors indicator in our data factor indicator are dichotomized variables the basic objectives of EFA are

1. explore the interrelationships between the observed variables

2. determine whether the interrelationships explained by small number of latent variables

The introduction and definition of EFA in this section were extracted from the book of Hardel and Simar (2003). P dimensional random vector x with mean \( \mu \) and covariance matrix \( \Sigma \) can be represent by matrix of factor loading \( Q_{pxk} \) and factor \( F_{kx1} \). The number of factor \( k \) should always be much smaller than \( p \).

\[
X_{(px1)} = Q_{(p \times k)} \cdot F_{(k \times 1)} + \mu_{(p \times 1)}
\] (1)

If \( \lambda_{k+1} = \ldots = \lambda_p = 0 \) (Eigen values of \( \Sigma \)), one can express \( X \) by the factor model 1. At that time, \( \Sigma \) will be a singular matrix. In practice this is rarely the case. Thus, the influences of the factors are often split into common (F) and specific (U) ones. U is a \( (p \times 1) \) matrix of the random specific factors which capture the individual variance of each component. The random vectors F and U are unobservable and uncorrelated.

\[
X_{(px1)} = Q_{(p \times k)} \cdot F_{(k \times 1)} + U_{(p \times 1)} + \mu_{(p \times 1)}
\] (2)

It is assumed that

- \( EF = 0 \)
- \( \text{var}(F) = I_k \)
- \( EU = 0 \)
- \( \text{cov}(U_i, U_j) = 0, i \neq j \)
- \( \text{cov}(F, U) = 0 \)

EFA is normally performed in four steps
1. Estimating the correlation matrix $\hat{R}$ between the variable one use person correlation coefficient if the data are metrical for ordinal data Kendall’s $\tau_b$, spearman’s rank correlation or polychoric correlations are applied.

2. Estimating the number of common factor $K$, $K \leq P$

3. Estimating the loadings matrix $\hat{Q}$ of the common factor there are different methods for computation of factor model principal (PC), principal Axis (PA), maximum likelihood (ML) and un weighted least square (ULS)

4. Rotation of factors loadings helps to interpret the factor easier, e.g. varmax, promax.

3 Estimating of common factors

The aim of factor analysis is to explain variation and covariance of multivariate data using fewer variables, the so called factors the unobserved factors are much more interesting than the observed variable themselves. How do the estimate procedure looks like

The estimation of factor model is based on the covariance or correlation matrix of data $x$ can be shown as follows After standardizing the observed variables the correlations of $x$ is computed with the similar form as the covariance matrix.

$$\Sigma = E(X - \mu)(X - \mu)^T$$

$$= E(QF + U)(QF + U)^T$$

$$= QE(FF^T)Q^T + E(UU^T)$$

$$= Q \text{var}(F)Q^T + \text{var}(U)$$

$$= QQ^T + \Psi$$ \hspace{1cm} (3)

The objective of FA is to obtained the loading $Q$ and the specific variance $\Psi$ the factor loading are not unique taking the advantage of the this uniqueness we get loading matrix multiplication by orthogonal matrix which can make interpretation of factors easier as well as more understandable. Factor loading matrix $Q$ gives the covariance between observed variables $x$ factor it is very meaning full to find another rotation which can shows the maximal correlation between factor and original variables.
4 Confirmatory factor analysis

Confirmatory factor analysis (CFA) is used to test whether the data fit a hypothesized measurement model proposed by a researcher. This hypothesized model is based on theory or previous study. The difference to EFA is that each variable is just loaded on one factor. Error terms contain the remaining influence of variables. The null hypothesis is that the covariance matrix of the observed variables is equal to the estimated covariance matrix.

\[ H_0: S = \Sigma(\hat{\theta}) \]

Where is \( \Sigma(\hat{\theta}) \) the estimated covariance matrix.

5 Principal Component analyses (PCA)

The objective of PCA is to reduce the dimension of multivariate data matrix X achieved through linear combinations. The principal component analysis (pc) chosen the highest variance of the data whose direction is the eigen vector corresponding to the highest eigen value \( \lambda_1 \) of variance matrix \( \Sigma \) orthogonal to the direction \( y_1 \) we find the second highest variance We centered the variable \( x \) to obtain a zero mean pc variables \( Y \).

\[ Y = r^T(X - \mu) \tag{4} \]

The variance of \( y \) will be equal to the eigen value \( \Lambda \) the component of the eigen vectors are the weights of the original variables in the pcs the principal component method in factor analysis can be done as follows.

- spectral decomposition of empirical covariance matrix
  \[ S = r\Lambda r^T \]
- approximation loadings \( \hat{\theta} = [\sqrt{\lambda_1 y_1} \ldots \sqrt{\lambda_k y_k} \] where \( k \) is number of factor
- estimating of specific variance by \( \hat{\psi} = S - \hat{Q}\hat{Q}^T \)

Residual matrix analytically achieved from principal component solution so that it is smaller than the sum of the falling eigen values.

\[ \sum_{ij} (S - \hat{Q}\hat{Q}^T - \hat{\psi})_{ij}^2 \leq \lambda_{k+1}^2 \ldots \lambda_p^2 \]
5.1 Number of extracted factor

Kaiser criterion, According to Kaiser factor should be extracted when their eigen values are bigger than one an eigen values of factor indicates variance of all variable which are explained by the factor.

5.2 Rotation of factors

Varmix rotation method proposed by Kaiser is orthogonal of factor which maximizes the variance.

6 Test of model fits

The $X^2$ test statistics goodness of fit indices such as RMSEA, TLI and CFI are used to evaluate to what extent a particular factor model explain the empirical data

7 Chi-square test

The $X^2$ test checks the hypothesis that the theororical covarianace matrix corresponds to the empirical covariance matrix the test statistics is $x^2$ distributed under the assumption of null hypothesis the null hypothesis will be rejected if the values of the test statistics is large.

$$X^2 = (n - 1)F(S, \Sigma(\hat{\theta}))$$  \hspace{1cm} (5)

Where $n$ is the number of observations, $F$ is the minimum of the discrepancy.

8 Root mean square Error of Approximation (RMSEA)

RMSEA is a measure for the model deviations per degree of freedom which ranges from 0 to 1 a value less than 0.06 indicates good model fit the values which are higher than 0.06 are indicative of bad model fit the model is acceptable when the values is greater than 1.

$$RMSEA = \sqrt{\frac{x^2}{df - 1}} \frac{1}{n-1}$$  \hspace{1cm} (6)
9 Akaike’s information criteria and the Bayesian information criterion BIC

In fitting latent models to data determining the number of latent remains a challenge to analysis in general, latent model processed by starting the most persimmons model and fitting successive models Akaike’s information criteria and the Bayesian information criterion BIC are widely used (Anderson, 1982). AIC is a measure of the goodness of fit of a model that considers the number of model parameter (p) being estimated in the model.

\[
AIC = -2 \ln L + 2p. \tag{7}
\]

AIC is an information criterion for ordering alternate models for data. The individual AIC values are not meaningful and much affected by sample size. Only those difference in AIC are interpretable as to the strength of evidence.

\[
\Delta_i = AIC_i - AIC_{min} \tag{8}
\]

Where \(AIC_{min}\) is the minimum of the possible \(AIC_i\) values some rules of thumbs are often useful in assessing the relative merits of model in the set BIC is a measure of the goodness of fit of the model that consider the number of parameter (p) and the number of observation (N).

\[
BIC = -2\ln L + (\ln N)p \tag{9}
\]

As with AIC the model with the smallest value BIC among all possible models are selected. The BIC applies larger penalties per parameter of \(\ln(N)\) than AIC, thus other factors being equal BIC tends to select Simpler model than AIC.

10 Fit Statistics Conditioning on Total Score

Instead of using ability estimates to divide examinees into subgroups, Orlando & Thissen (2000) partitioned examinees into subgroups based on the observed total test score. They proposed new fit statistics (\(s - X^2\)) based on traditional \(X^2\) the \(s - X^2\) statistics is defined as
\[ s - X_i^2 = \sum_{k=1}^{n-1} N_k \frac{O_{ik} - E_{ik}}{E_{ik}(1 - E_{ik})} \]

(10)

where \( N_k \) is the number of examinees at score subgroup \( k \).

\( O_{ik} \) is the observed proportion of correct response to item \( i \) by subgroup \( k \) and

\( E_{ik} \) is the expected proportion of correct response to item \( i \) by subgroup \( k \).

11 Good ness of test statistics (GOF) for Assessing over all fit.

Maydeu –Oliver and oe (2005,2006) proposed a family of GOF statistics, \( M_r \) that provide a unified frame work for limited information and full information GOF statistics this family can be written as

\[ M_r = N \hat{e}_r C \hat{e}_r \]

(11)

Where \( \hat{e}_r \) are the residual proportions up to order \( r \) and

\[ C = \Gamma r^{-1} - \Gamma r^{-1} (\Delta_r^{\prime} \Gamma r^{-1} \Delta_r)^{-1} \Delta_r \Gamma r^{-1} \]

Hence \( \Gamma r \) represent the asymptotic covariance matrix of the residual proportions up to order \( r \) and \( \Delta_r \) is a matrix of derivatives of the marginal probabilites up to order \( r \) with to the model parameter two member of this familys for instanse \( M_2 \) and \( M_n \), in \( M_2 \) only univariate and bivariate residuals up to order \( n \), the number of variables ,are used .

The asymptotic distribution of any statistic of the \( M_r \) family is chi-square with degree of freedom(df)=number of residual used-q for the ch-square approximation to \( M_r \) be calculate the expected frequencies of min (2r,n) marginal need to be large.Thus,for \( M_n \) expected cell frequencies need to be large ,but for \( M_2 \),where \( r=2 \) only expected frequencies for sets of min (2r,n)=4 variable need to be large as result ,when only lower order margins are used ,the p-values are accurate.

10 Information function

The information function (IF) for the GR model derived by Samejlima(1996) is given in the following equation. The IF point out how many standard error of the trait estimate are needed to
equal one unit on the proficiency scale when more standard error are needed to equal one unit on the $\theta$ scale. The standard error is smaller indicating that the measurement instrument is sensitive enough to detect relatively small differences in $\theta$ (Lord, 1980 and Hambleton and Swaminathan, 1985). The information for scale item in the GR model is weighted sum of the information from each of the response alternatives is the information for a scale item in GR model is a weighted sum of the information from each of the response alternative if the information plot is.

$$I(\theta, u_i) = \frac{\sum_{k=1}^{n+1}[\alpha_i p_{i,k-1}(\theta) q_{i,k-1}(\theta) - \alpha_i p_{i,k} q_{i,k}(\theta)]^2}{p_{i,k-1}(\theta) - p_{i,k}(\theta)}$$ (7)

Where $p_{i,k}(\theta_j) = p^*(u_{ij} = k|\theta_j)$ is a cumulative category response function of $k$ steps of the item and $q_{i,k}(\theta_j) = 1 - p_{i,k}^*(\theta_j)$

11 Differential item functioning (DIF)

DIF is an important issue in large scale standardized testing it refers to the unexpected different interims of performance among groups of equality proficient examiner (Lord, 1980), thus is presence would threaten the validity of inference drawn from a test a variety of method for assessing DIF have been developed Pontenaza and Dorans (1995) overviewed the method for finding DIF items and classified them on the basis of whether they define DIF with respect to an observed variable or a latent variable and whether or not a parametric form describe the relationship between item response function and the latent trait (Hambleton and Swaminathan, 1995). Item response function (IRF) describes the item response probability of getting an items correctly given the latent trait level if an items does not display DIF then its IRF should be the same for all groups under consideration on the other hand if an item display DIF, the IRFs will be different across groups More technically DIF is said to occur whenever the IRF, $p_j(\theta)$, of getting correct response differ for the two groups For item $j$ the null DIF hypothesis is

$$H_0: p_{g1,j}(\theta) = p_{g2,j}(\theta), \text{for all } \theta.$$  

The IRF is characterized by items parameter (e.g. $\alpha, \beta$). DIF falls into two categories: uniform DIF and uniform DIF in the two parameter logistic model when $\alpha_{g1,j} = \alpha_{g2,j}$ but $\beta_{g1,j} \neq \beta_{g2,j}$ the two IRFs are parallel and there is a location shift due to different Groups membership the uniform DIF against the same groups for all ability regardless of whether the magnitude of
DIF is constant or not constant as ability level varies uniform DIF occurs when \( p_{g1}(\theta) \geq p_{g2}(\theta) \) or \( p_{g1}(\theta) \leq p_{g2}(\theta) \) for all \( \theta \).

The amount and the direction of DIF can vary at different ability level. non uniform DIF occurs when the discrimination parameter differ \( a_{g1,j} \neq a_{g2,j} \) across groups, therefor the probability of getting an items right for the two groups change sign over the ability range. In IRT terms, non uniform DIF is indicated by two cross item response function while uniform DIF is displayed by two non-crossing item response function Thissen et al (1988) introduce model comparison measure which is implemented by likelihood ratio test to detect DIF they used Marginal maximum likelihood, the likelihood for maximizing is marginalized with respect to \( \theta \) and has the form.

\[
L(\alpha, \beta) = \prod_{i=1}^{N} \int \prod_{j=1}^{I} P_j(\theta_i)^{y_ij} (1 - p_j)(\theta)^{(1-y_ij)} dF(\theta_i) 
\]

The likelihood ratio test involves the comparison of two models, a reduce model under \( H_0 \) and full model under \( H_1 \) the test statistics is.

\[
G^2 = -2 \log \left[ \frac{\max_{w_0} L(\alpha, \beta)}{\max_{w_1} L(\alpha, \beta)} \right] 
\]

where \( w_0 \) is the parameter space under \( H_0 \) and \( w_1 \) is the parameter space \( H_1 \), its p-value is obtained by chi-square distribution DIF analysis based on IRT model examinees whether IRFs perform differently for manifest variable such as race or gender by allowing the items parameter to vary from each group.

DIF could be viewed as a misfit of the IRT model because the existence of DIF implies that there is another factor influence the items response probability (Glas, 2001; Thissen et al, 1993). Also using background variables could be use full way to investigate why an observation occurs. For example differences of items response function between races could be explained by social-economic status (Glass, 2001; Rogress et al., 1999).