ON TESTING AND ESTIMATION IN THE ECONOMIC MEASUREMENT, WHEN USING ITEM RESPONSE MODELS

RĂILEANU SZELES MONICA
Department of Finance, Accounting and Economic Theory
Transylvania University of Brasov, Romania and CEPS/ INSTEAD, Luxembourg
29, B-dul Eroilor, Brasov, 2200
Romania
monica.szeles@unitbv.ro

Abstract: - This paper applies the one and two parameter- item response models (IRT) with a probit link to derive a scale of financial precariousness. A strong focus is on the item testing and selection, which is sometimes ignored in the IRT literature. In order to test whether the items fulfil or not the IRT fundamental assumptions, a number of tests are applied. The Mokken Scale is the most comprehensive one and provides the most complex output. Also, the paper highlights the advantages of using the IRT instead of traditional models when deriving latent measures from a set of observable binary indicators. This is often the case in the economic and social fields where the IRT is rarely used. Using both IRT models and some testing procedures, we derive a scale of four items and estimate item parameters. In the empirical part, the IRT models are applied on data from the Luxembourg socioeconomic panel (PSELL-3, wave 2006) using the modules GLLAMM (for running the IRT models) and msp/loevH (for testing the items) in the statistical package STATA. The software and modules allowing the application of the IRT models and its underlying tests are also examined. At the end, the paper leads to both empirical and methodological conclusions.

Key-Words: - Item Response Theory, Scale of deprivation, Tests, Difficulty, Discrimination, Parameter, Score.

1. Introduction

The item response theory (IRT), often referred to as latent trait theory or strong true score theory describes the application of mathematical models to data for measuring abilities, attitudes, scores or other latent variables. Although the IRT models have been used in psychometrics and educational testing for over a half a century as a traditional methodology [11], a small number of papers have applied them to the analysis of social and economic measures [2], [3], [5], [6], [7]. Recently, IRT models has been applied in the health economics and market research (especially the Rasch model).

IRT is also called a modern test theory because is derived from the classical test theory and belongs to the wide field of the item analysis. The IRT models can be used whenever we want to derive a latent summarizing measure from a set of observable dichotomous or polytomous items (indicators). This is often the case in economics, business or social sciences, where the traditional approach is to simply aggregate variables in order to represent socio-economic measures. Despite of its large use in the psychological and educational fields, the application of the IRT models to the economic and social fields is rather weak. Social and economic extensions of the IRT could be: the analysis of the financial position of a company, based on a set of indicators on the financial performance (finance) and the estimation of social and economic constructs, such as deprivation and wealth indices (welfare and poverty literature).

A new challenge in the economic and social measurement is to apply new techniques and models which are taken from other fields and disciplines and then to compare these results with those obtained by using traditional methodologies. For instance, interdisciplinary measurement techniques such as the fuzzy models [4], [12], [15], [17] and the spatial modelling [12] have been recently extended to the economic and social fields. The IRT models can be also seen as an innovative methodology.

In comparison with traditional methodologies (e.g. classical test theory), IRT brings two main advantages. First, it relates characteristics of items (item parameters) and characteristics of individuals (latent measures) to the probability of providing a particular response. Second, items and persons are represented on the same metric scale, which can be viewed as an optimal scale design. Also, the IRT allows dealing

with missing values, without being necessary to drop them in advance.

In this paper, the IRT are applied to the measurement of financial precariousness, which is a fundamental dimension of deprivation. At a conceptual level, deprivation describes a situation where a person lacks several goods and services. It therefore has a multidimensional and relative nature. Financial precariousness is unidimensional and can be described either by one indicator such as income or by a set of financial indicators [14].

Because deprivation is a latent construct, its measurement requires aggregation of items into an index. The traditional approach to deprivation measurement is the weighting or not-weighting summing up of items (sum- score approach). In this paper we analyze an alternative methodology, which allows not only deriving person parameters (e.g. index or score), but also item parameters (difficulty and discrimination of items).

The main focus of the paper is to explore the potential that the IRT models have in the economic and social measurement. The probit one and two parameter- IRT models are particularly investigated here and the analysis regards not only the comparison of the two IRT models, but also the testing and selection of items. Lately, the development of specific IRT software, such as BILOG, MULTILOG, as well as the specific modules that have been especially created for the IRT analysis in the statistical packages SAS and STATA, allows testing the items and estimating the IRT parameters. This is much easier to do in the context of a static analysis. When moving from a cross-sectional to a longitudinal approach, software issues may arise.

2. Econometric model and software challenges

The IRT models describe the relationship between item responses and the latent summarizing variable, through a link function. The most important advantage that it brings in comparison with the classical test theory is that it relates the characteristics of item (items parameters) and characteristics of individuals (person parameters), both being represented on the same metric scale.

According to the number of parameters required in order to model the responses to each item, the IRT models are 1 parameter IRT, 2 parameter- IRT and 3 or 4 parameter- IRT. Upon the link function, there are two types of IRT models: the normal IRT models (normal ogive models), which are based on the cumulative normal probability distribution function while the logistic models (logistic ogive models) are based on the logistic function.

In the general one parameter IRT model, the modelling of the observed dichotomous items of deprivation allows estimating only a single item parameter and a person parameter. As the deprivation items are dichotomous, they take in this model either the value *0*, if the individual doesn't experience that symptom of deprivation, or the value *1* if he does.

The item parameter is called item difficulty parameter and represents the intrinsic meaning of that indicator [3]. The higher the difficulty parameter is, the more difficult to endorse a positive answer. The person parameter is the individual score of deprivation. According to the IRT propriety of invariance, the item and person parameters are invariant depending neither on the subset of items used, nor on the distribution of the latent trait in the population of respondents.

In the equation below, β_i is the difficulty parameter for the item i, D_j^* is the latent score of deprivation for the individual j and V_{ij}^* is the latent indicator of deprivation for the item i and individual j. ε_{ij} is a normally distributed error term with mean zero and fixed variance.

$$V_{ij}^* = \beta_i + D_j^* + \varepsilon_{ij} \quad (1)$$

$$V_{ij} = 1 \ \ if \ \ V_{ij}^* > 0 \ \ and \ \ V_{ij} = 0 \ \ otherwise \quad (2)$$

If we treat D_j^* as random individual effects, then the standard maximum likelihood provides estimates of both the parameter β_i and the deprivation score D_j^* [1]. Although the deprivation scores can be estimated by using empirical Bayes methods, this could raise some methodological problems. First, an accurate prediction is conditioned by a large number of deprivation indicators. But if a large number of deprivation indicators will be included in the analysis, it will become difficult to estimate the model. Second, as the estimates of the EB predictions and the sum-score method are closely related, the results provided by the IRT model could not reveal significant new patterns or a new scale of deprivation.

The one parameter- IRT model where the error term has a logistic distribution is known as the Rasch model. In this particular case, the total score is proved to be a sufficient statistic of the individual ability parameter. The Rasch model has a unique measurement propriety which distinguishes it from other models used to model person's responses to items. This means that the simple aggregation of the indicators respecting the Rasch model assumptions

gives the deprivation score [13]. The most difficult part here is to find those indicators which fulfil this restrictive condition. It is not only the Rash model, but the one parameter IRT model in general that imposes strong assumptions, like the equi-correlation between any pair of deprivation items. In the two-parameter IRT model, this strong assumption is not required. Even though the Rasch model is a very restrictive one as the items must fulfil a set of conditions, it has a nice interpretation. On short, it allows depicting a set of items which can be aggregated into a scale or index.

For some reasons, the use of the two parameter IRT model carries some advantages, in comparison with the one parameter IRT model. The most important advantage is that with the two parameter- IRT, we can estimate not only one, but two item parameters or even more: the item difficulty parameter (β_i) and item discrimination parameter (λ_i). The discrimination parameter indicates how well an item discriminates along the scale of deprivation continuum. The higher the discrimination parameter is, the more desirable the question.

$$V_{ij}^{*} = \beta_i + D_j^{*}\lambda_i + \varepsilon_{ij} \quad (3)$$

$$V_{ij} = 1 \;\; if \;\; V_{ij}^{*} > 0 \;\; and \;\; V_{ij} = 0 \;\; otherwise \quad (4)$$

The disadvantage is that in the two parameter- IRT model, the propriety of sufficiency of the score on the latent trait does not hold anymore, because a change in the latent score of deprivation does not equally affect the items of deprivation.

There are two types of software which allow estimating the IRT models. Special software have been designed for the IRT models, such as BILOG, LOGIMO, MSP, MULTILOG, PARELLA, PARSCALE, TESTFACT and XCALIBRE. But also statistical software such as STATA and SAS, which are generally widely used in the economic and social sciences, can deal with these models. Most of the software above estimate item and person parameters, test and analyze items by local or global tests, draw IRT curves (such as item characteristic curves, information curves) and can deal with any number of subtests/ subscales. But the number of IRT tools provided by the statistical packages STATA and SAS are limited, compared to the specific IRT software.

In STATA, the IRT models can be applied using the module GLLAMM. GLLAMM estimates generalized linear latent and mixed models by maximum likelihood [9]. As the item response models belong to this class of models, the paper also investigates the potential of the STATA program for this kind of analysis.

3. On the importance of item testing

In order to be included into a scale, the items must fulfil a set of assumptions which attest whether they are reliable and describe a single predominant trait. Whenever we aggregate variables into indexes or scales, a number of tests apply to check the appropriateness of items. Most of them are traditional tests, such as the factor analysis**,** Loevinger's scalability coefficient H and Alpha Cronbach statistic. When using the IRT as a measurement technique, a set of tests must be designed and analyzed in the framework of this theory. The test design is a process which consists of extracting a sub- set of items from a big collection of items (item pools).

All IRT models rely on a set of fundamental hypotheses:

- Unidimensionality of latent trait: The first central assumption in the IRT is that the items measure just one latent trait. This hypothesis implies that a single dominant trait gives the probability of item endorsement, although in practice it is usually assumed that minor violations don't make such difference.
- Local independence: The second central assumption is the local independence. According to this assumption, after controlling for dominant factors, item pairs should not be associated. The local independence relates to the unidimensionality in the sense that no other characteristic of the individual influences the response probabilities.
- Monotonicity: The third assumption arises from the distinction between parametric and non-parametric IRT. In parametric IRT, the probability to responding positively to an item is a continuous function, with parameters indicating the location, the slope and the asymptotic values. In nonparametric IRT, where order restrictions are imposed on the IRT functions, the main assumption is the monotonicity. Due to this assumption, the slope of the IRT function looks like being non-decreasing.
- Invariance : IRT item and person parameters should be invariant, depending neither on the set of items used, nor on the distribution of the latent variable in the population. The assertion that the parameter estimates will not depend on the sample used holds only when a set of constraints are placed on the sample distribution

and on other parameter values. The constraints compress and expand the latent variable which loses the linear form and becomes a local description. But this is the trade-off to always match the empirical data. The propriety of invariance makes the major difference between IRT and classical test theory.

In practice, the unidimensionality is the most important assumption under all IRT models. In literature, the most popular methods are determining the number of eigenvalues greater than 1, examining the scree plots and considering the ration of the first eigenvalue to the second. The Modified Parallel Analysis (MPA) and DIMTEST [16] are software tools specifically developed for IRT analyses. DIMTEST assesses the degree to which two subsets follow the same factor pattern, but requires around 20 items for reliable results. MPA determines whether specific real data is sufficiently unidimensional for IRT analyses. This procedure is based on a comparison between the eigenvalues of a real dataset and a unidimensional synthetic dataset which considers the parameters obtained from the real data.

Another procedure which is usually applied whenever we construct scales is the reliability analysis. This is because, in order to summarize the items into a scale, we first need to test whether they have a "common" part. The reliability of the deprivation scales derived in the empirical section is then assessed by the Cronbach alpha statistic, which can be interpreted as an indicator of internal consistency. In the classical test theory, *alpha* is an unbiased estimator of reliability only under a certain condition. The components may have different means and variances but they must have equal covariance, which requests that they have one common factor in a factor analysis. In general, *alpha* increases when the correlation between the items also increases, and it is mostly used when the items measure different substantive areas within a single latent construct.

The selection of items related to the value of the Cronbach *alpha* is thus essential, as they should reflect a high internal consistency within the deprivation scale.

Even though the item characteristic curve (ICC) cannot be considered as a test itself, it is a useful graphical tool to analyze and see the appropriateness of items. The ICC describes the relationship between the underlying deprivation score and the response to each item of deprivation scale. In fact the ICC is a two-dimensional scatter plot of deprivation scores by item-response probability, depicting the item response that would be expected from an individual located at any given point on the underlying construct. Therefore, for each item of scale we have one ICC. The distribution of deprivation scores don't need to follow a particular form (e.g., a normal distribution). In the one parameter- IRT all items exhibit ICCs having the same shape, because we assume in this model equal discrimination power for all items. In this case, the difference between the ICCs of a particular scale is in the level of the deprivation score that is associated with a given observed probability of a keyed response.

The item difficulty is a location index describing where the item functions along the ability scale. The higher the "difficulty" parameter, the more difficult to endorse a positive answer (to possess that item) is. The steepness of the ICC in its middle section reflects the discrimination power of items. The steeper the curve, the better the item can discriminate because the probability of a correct response at low deprivation scores is not the same as it is at high deprivation scores. The flatter the curve, the less the items discriminate since the probabilities of correct response at low and high deprivation rates are nearly the same.

Testing the fulfilment of the hypothesis above allows indirectly determining a set of items which respect the IRT assumptions. This can be considered as a selection procedure. Some authors give a high importance to this question [5], [8], when others do not put much emphasis on it [3].

Over time, software tools have been developed to incorporate all tests into one global test. For instance, in the statistical package STATA, the MSP module implements the Mokken Scale Procedure, which constructs sub-scales based on the H Loevinger's coefficients. Another module in STATA is LoevH. This one computes the Loevinger H coefficients and the Guttman errors for each pair of items, between one given item and all the others of a scale or among all the possible pairs of items of a scale [8]. In other worlds, the Mokken Scale Proceedure is an item selection proceedure, meaning that it detects from a set of items the number of scale and the composition of each scale.

4. Analyzing deprivation in Luxembourg by using IRT models

As was introduced in the sections above, the IRT methods can be applied whenever we want to derive a latent score from a set of dichotomous or polytomous variables.

In the empirical part we apply the probit one and two parameter- IRT on data from the national panel of Luxembourg, called Socio-Economic Panel

"Liewen zu Lëtzebuerg" (PSELL), at a cross-sectional level.

The analysis uses data from the wave 2004 of PSELL3, which runs from 2003 to 2006. PSELL dataset covers the residents who live in the Grand Duchy and who are protected by social security. The initial sample included 6110 persons and 2012 households, covering 97% of the population living in Luxembourg. Since 1985, PSELL was conducted every year, through three subsequent panels: PSELL-I (1985-1994), PSELL-II (1995-2002) and EU-SILC/PSELL3, which was launched in 2003 and is currently running.

The financial precariousness is defined here as a latent measure which initially summarizes five observable items describing the domain of financial deprivation. If the factor analysis confirms that the items describe not only a domain, but a single dimension, then IRT models may be used.

The variables of financial precariousness and the underlying rates of deprivation are:

- Enough money to eat meat each second day: 2.48%
- Being unable to save money: 52.92%
- Afford one holiday away from home: 14.01%
- Ability to face unforeseen emergencies: 24.12%
- Making ends meet: 23.71%

All variables of our analysis are dichotomous, as it is required by the IRT models that we use here[1]. The value *1* suggests the financial precariousness and the value *0* suggests the absence of financial precariousness.

4.1. Testing and selecting the items of deprivation scale

As item testing and selection is an important part of the IRT methodology, we consider it as being the first step of our analysis. This can be seen as an innovative contribution to the analysis of deprivation by one and two parameter- IRT models. Other authors have either applied tests in order to measure deprivation by the Rasch model [5], or have simply applied both the one and two parameter- IRT, without testing the items [3]. The tests which are usually used as a part of the IRT methodology are local or global tests, testing the items of scale or the whole model and its goodness of fit. In the case of the Rasch model, which is the most restrictive one parameter- IRT model, specific tests have been created to test whether the items meet this set of particular requirements. But when using other IRT models, the literature is not very clear and concise about what tests to apply. Independent on their complexity, they all measure to what degree a set of items meets the IRT fundamental assumptions, as presented in a section above.

In order to test whether the items of financial precariousness fulfil the IRT assumptions[2], the Mokken scale [8], the confirmatory factor analysis and the reliability analysis are applied.

The most important IRT assumption is the unidimensionality. In case that two or more prominent dimensions are discovered, multidimensional IRT models should be applied. The easiest way to assess the unidimensionality of a scale is by the factor analysis.

For the "financial precariousness" scale only one factor is retained. This one explains more than 50% of the total variance and represents the dimension of "financial precariousness". The first eigenvalues is 1.86 and the second is 0.05, while the rest are negative[3]. Although we identify at this step the unidimensionality of scale, we also see that the item "savings" is negatively related to the main factor, which suggests dropping it.

The propriety of monotony can be assessed by drawing the item characteristic curves. In fact, the graphical representation of the ICC curves in the next section reflects the propriety of "double monotonicity": for each level of deprivation, the probability of not being deprived decreases with item difficulty and for each item difficulty, the probability of not being deprived increases with the level of deprivation.

The invariance of parameters is the third propriety of IRT. According to IRT, the parameters of ICC must be invariant across population. It means that if one picks different samples and estimates the ICCs, he should get the same values of parameters and the same ICCs. This should happen because if he has part of the curve, he can recover

[1] Not only the dichotomous items can be modelled by an IRT model, but also other types of data (polytomous data, for instance). Continuous ordered data can be modelled by the continuous response model (Samejima model) or continuous Rash model, discrete nominal data by the nominal response model and discrete ordered data by the acceleration model (Samejima model), polytomous Rash models or generalized partial credit models (Muraki model). These models can be also adapted to the measurement of deprivation, when items are defined as polytomous variables.

[2] Unidimensionality, monotonicity and local independence are the fundamental assumptions of the IRT models [8].

[3] Unidimensionality is confirmed by factor analysis when the enginevalue of the first factor is at least three times higher than the others.

the rest of it. In order to get accurate estimates, population should cover a high part of the curve. Also, a big variation in the population deprivation scores is useful in estimating item parameters. According to our estimates, the variation of the latent measure is 1.53 (0.32 standard error) in the two parameter- IRT model and 3.71 (0.22 standard error) in the one parameter- IRT model. This variance suggests a good representation of population for an accurate estimation of item parameters.

To investigate the degree to which the propriety of invariance holds for the item difficulty parameters, the initial sample was divided into four sub-samples. The first two sub-samples results by splitting the initial population into two randomly equivalent sub-groups. Differences between the item difficulty parameters between sub-groups appear as not significant. Other two sub-samples are generated by splitting the initial population upon gender. Same conclusion arises as for the first split, and this leads to the conclusion that the assumption of IRT item parameters invariance holds.

The last procedure that we apply here is the Mokken scale, even though this is specially conceived for non-parametric IRT models. In this paper we use it to test not only just one assumption, but the whole set of IRT assumptions. In STATA, the Mokken Scale Procedure (msp) is a module which displays for each sub-scale, the Loevinger H index, and the Loevinger Hj indexes for each item. In some cases, both loevH and msp module disply the Loevinger H indexes, which makes them almost similar. But when applying them to the scale of financial precariousness, the item „savings" makes the difference. With the msp proceedure, the item „savings" is not included in any scale, while with the loevH proceedure, which includes all items into the scale, the Loevinger index is 0.11, which is a very low value. This is another hint to exclude this item from the scale of financial precariousness.

The Loevinger index for the new scale of 4 items is 0.701 which indicates that they may form a good scale according to the IRT assumptions. Also, the analysis provides the rankings of items upon their difficulty, as it is presented below:

Difficulty of items:

- Enough money to eat meat each second day: 0.97
- (Being unable to save money: 0.47)
- Can afford one holiday away from home: 0.86
- Making ends meet: 0.76
- Ability to face unforeseen emergencies: 0

The reliability of the financial precariousness scale can be tested by the Cronbach alpha statistic. In general, *alpha* increases when the correlation betwee items also increases, and it is mostly used when the items measure different substantive areas within a single latent construct. The selection of items upon the value of *alpha* is essential, as they should reflect a high internal consistency within the deprivation scale.

For the scale of 5 items the alpha statistic is 0.737, while for the scale of the 4 items which are finally selected to define financial precariousness alpha is 0.739. This is in line with the value of 0.653 reported by Cappellari and Jenkins [3]. It indicates a strong scale and a high internal consistency.

As shown in this section, the item testing and selection is an important part of the IRT models. In our case, all tests suggest eliminating the item "savings" from the scale of financial precariousness, even though it has a financial nature. It is not only the IRT assumptions that recommend the rejection of this item, but also the proportion of people defined as deprived upon this criterion. For the item "savings" this rate is higher than 50% meaning that the inability to save money is prominent in society. Considering this evidence, the ability to save money cannot be viewed as a symptom of deprivation.

4.2. Estimates from the IRT models

In this section we apply the one and two parameter- IRT models to estimate two item parameters: difficulty and discrimination power. This allows ranking the items of scale and also comparing the estimates generated by the two models.

The two IRT models are estimated in STATA, using the module GLLAMM [19]. The post-estimates allow deriving the item characteristic curves for the items of financial precariousness.

As shown in the table 1, the estimates of the difficulty parameters show the same ranking of items in both models used here. The most "difficult" item is "ability to face unforeseen emergencies" and the "easiest" one is "enough money to eat meat each second day". "Afford one holiday away from home" is the most discriminant item and "enough money to eat meat each second day" is the least discriminant. The estimates and the likelihood ratio test suggest that all the five items have different discrimination powers.

In terms of IRT, for any level of individual deprivation score the probability of reporting a symptom of deprivation is the lowest for the item having "enough money to eat meat each second day", and the highest for "ability to face unforeseen

emergencies". In terms of Rasch model, this means that the probability that an individual who does not have enough money to eat meat each second day to also be unable to face unforeseen emergencies is higher than 0.5. The log likelihood values indicate a small improvement when moving from the one parameter IRT to the two parameter- IRT.

The item ranking provided by the IRT models is similar with that given by the ranking of their rates of deprivation and also with the ranking upon difficulty in the Mokken Scale model. Cappellari and Jenkins [3] report the same similarity (excepting the Mokken Scale), but without considering testing and selection issues. We note at this point that the empirical analysis does not always lead to the similarity between the rankings of items in the two IRT models, as in this case.

Table 1
Estimates of item parameters from probit IRT models

Variables	One parameter **IRT**		Two parameters **IRT**			
	β_i	St.err.	β_i	St.err.	λ_i	St.err.
Enough money to eat meat each second day	-4.23	0.1143	-3.15	0.1995	1(fixed)	-
Ability to face unforeseen emergencies	-1.56	0.0636	-1.47	0.0745	1.42	0.1721
Afford one holiday away from home	-2.41	0.0761	-3.12	0.5100	2.10	0.4141
Making ends meet	-1.61	0.0643	-1.65	0.0901	1.60	0.1964
Log Likelihood	-4844.0328		-4831.2311			

Notes. β_i is the item difficulty parameter and λ_i is the item discrimination parameter.
All coefficients are significant at 1 per cent level.
Likelihood ratio test of the two parameter- IRT model versus the one parameter IRT model: LR chi2(3) = 25.60, Prob > chi2 = 0.0000.

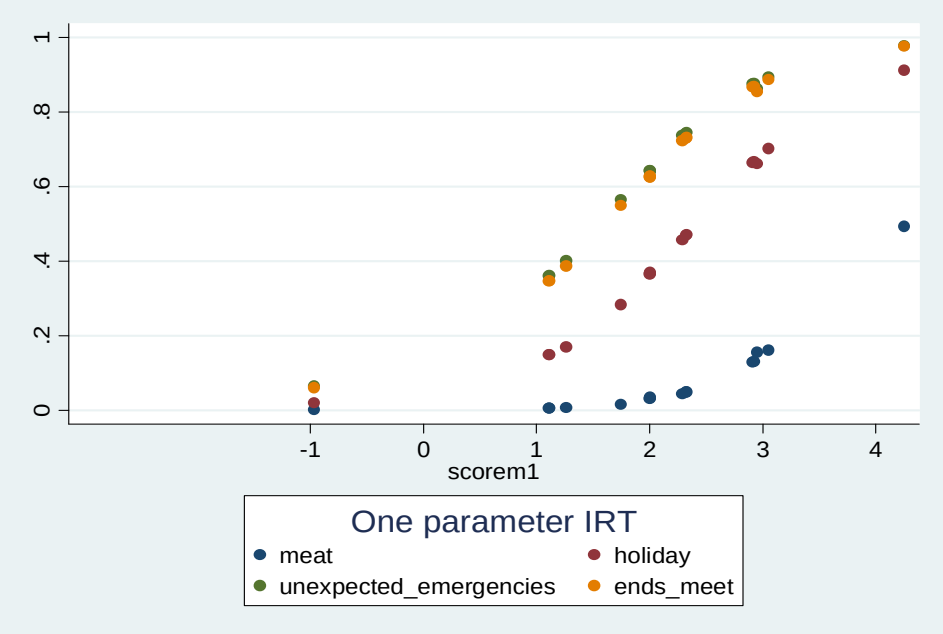
The item characteristic curves provide a nice interpretation of results. In order to draw the ICCs for the scale of financial precariousness, we set the scale by considering the mean of population equal to zero and the population standard deviation to one. Therefore, in the graph, the value 1 on the horizontal axis corresponds to 1 standard deviation above the mean and the value -1 to one standard deviation below the mean. This means that the probability of being deprived upon a certain item of deprivation increases as the score of deprivation increases.

In general, when drawing ICCs a first question to be addressed is whether the assumed curve is reasonable. For deprivation, such a relation as below appears reasonable. One problem that may arise concerns the borders of the deprivation score distribution. Data from large samples are in general necessary to make the curve behave properly at the borders of the ICCs graph because normally samples are small at the extremes of the deprivation score distribution.

The difficulty parameter is the most central parameter and it sets the location of the inflection point of the ICC on the horizontal axis. It shifts the curve from the left to right as the item becomes more and more difficult. At this point, the ICCs reveal exactly the same evidence as the IRT estimates generated by the GLLAMM procedure. The easiest item is "meat" and the most difficult one is "unforeseen emergencies".

The ICC functions of "ability to face unforeseen emergencies" and "making ends meet" are almost identical in both models, exactly as their rates of deprivation. This similarity is not good, suggesting that one may replace one of these variables with a new one, in order to have a broader representation of financial precariousness. This is because ideally the scale should include items having different levels of difficulty.

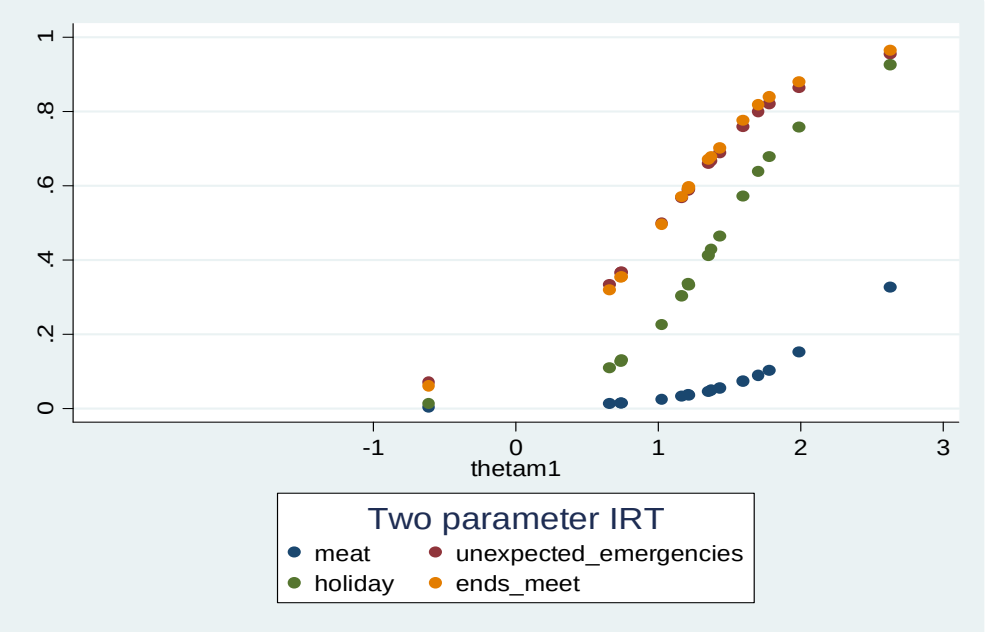
Figure 1: Item characteristic curves in the one parameter- IRT model



Notes. On the horizontal axis, scorem1 represents the score of deprivation and the vertical axis represents the observed item response.

The ICC analysis in the two parameter- IRT models also suggest that the most desirable item is holiday because it has the highest discrimination parameter. The ICC of this item is quite steep in the middle, where the probability of reporting a lack changes very rapidly as deprivation score increases.

Figure 2: Item characteristic curves in the two parameter- IRT



Notes. On the horizontal axis, thetam1 represents the score of deprivation and the vertical axis represents the observed item response.

As shown above in the figure 1 and figure 2, one difference between the ICCs representations in the one and two parameter- IRT relies on the item "holiday". In the two- parameter IRT model,

the ICC of the item "holiday" becomes very close to the ICCs of "ends_meet" and "unexpected_emergencies", at the bottom of the deprivation scores distribution. But this is not the same in the one- parameter IRT model. The explanation is that in the two parameter- IRT model the item having the highest discrimination parameter is "holiday" and this one discriminates better at the border of distribution.

The ICCs reveals another interesting interpretation. At high scores of deprivation, the probability to be deprived upon "holiday", "ends_ meet" and "unforeseen_emergencies" is high, but probability to face deprivation upon the item "meat" is low. This is because the item "meat" has the lowest difficulty and also the lowest discrimination power.

Even though in this paper we have used the probit link, other functions can be used as well. The most popular IRT models rely on the logit models (as is the case of the Rasch model). The option of choosing the link function can be seen as an advantage.

Even though issues related to the calculation of deprivation scores for different samples of population and analysis of determinants are not discussed here, the gllamm package for Stata can deal with them.

5. Conclusions

Lately, different methodologies have been advanced to summarize and to explain a number of deprivation indicators [10]. The IRT models provide an alternative framework to measure and analyze deprivation, by calculating not only scores or indexes (which are known in the IRT as person parameters), but also item parameters (difficulty and discrimination parameters).

This paper gives insights to the huge potential that the broad methodological framework of IRT has for the field of economic and social measurement. Whenever we want to aggregate a number of dichotomous indicators into a scale and/ or index, in order to define and measure a latent construct, the IRT may be used. There are a number of advantages that arise when using the IRT instead of traditional models in measuring latent phenomena.

The item response modelling is similar with the generalized latent variable modelling (multilevel, longitudinal and structural equation models) in the sense that technologies developed in one area can be exploited in another [19]. The IRT models are factor analytic measurement models for non-normal data in the exponential family. This means that the IRT can be thought of as generalized linear models combined with factor analytic models. For instance, the two parameter- logistic IRT model is equivalent to the Spearman factor model combined with logistic regression.

Additionally, the IRT provides a more complex output, including item testing, graphical representation and identification of determinants. In the economic and social field, the use of IRT is very new and here the empirical evidence is rather weak. When applied to the economic and social measurement, the importance of item testing has been rarely considered and this was only when using the Rasch model.

In this paper it was shown that the testing and selection of items is a very important step that should be always considered before running the IRT models. This stage allows deriving a strong, reliable scale, meeting the IRT fundamental assumptions. Even though the IRT literature does not provide a structured range of tests, traditional or specific tests may be also used. The Mokken Scale represents the most complex test which may replace the other tests that we have additionally used here. When applying the Mokken Scale in order to select the items of financial precariousness scale, we get that the msp procedure in STATA might be more useful than the loevH.

In our research, the empirical analysis leads to the same ranking of items upon their difficulty in the one and two- parameter IRT. This is in line with Capellari and Jenkins [3].

Apparently, the two IRT models seem to be similar, but in fact they are based on different assumptions and provide different kind of results. For instance, in the one parameter- IRT all items have the same level of discrimination, while in the two parameters, they are not equal. The simple analysis of items upon their rates of deprivation and the Mokken Scale analysis lead more or less to the same conclusions and ranking of items. Despite this, we prefer the IRT framework because this one is more complex and provides many tools for analysis, such as the ICCs. Also, it incorporates the analysis of determinants as a MIMIC model (Multiple Indicators- Multiple Causes), by supplementing the basic econometric model with a structural equation.

Lately, the development of specific IRT software as well as the creation of special modules in STATA or SAS will contribute to its adoption as an interdisciplinary measurement tool. Our paper has used the module GLLAMM in STATA, which provides interesting features for the IRT analysis. Also the SAS statistical package allows estimating the IRT parameters with the module nlmixed [19].

Other software such as Mplus, R, LatentGold provide similar facilities, though often based on different principles. No matter what software we use to estimate the IRT parameters, data must be formatted properly for that specific software.

At the end, we highlight the innovative contribution of the paper to the measurement of deprivation by IRT models. This consists of testing the items of scale and the scale itself. Without applying testing procedures, wrong items can be included into the model. In this case, the GLLAMM module or other module used for estimation cannot reject them. In the empirical part of our paper we have shown that the item "savings" should not be included into the scale of financial precariousness, upon testing procedures.

In this light, the selection of the items of financial precariousness, the measurement of financial precariousness and the analysis of its causes may be concentrated into a single IRT model.

In economy, as well as in other fields, the IRT proprieties make this model very useful for decision making purposes. For instance, it helps authorities in selecting the most representative basket of items for the definition and analysis of different symptoms of deprivation, such as financial precariousness. This is mainly due to the propriety of invariance, which does not hold anymore in other approaches, as the classical test theory for instance. With this propriety, the decisions we take are the same, being independent on items and population sample.

Although this paper touches a small part of the IRT methodology, it claims for its adoption into the economic and social fields. As the IRT is here in an incipient stage, it can be extended in many directions, at both cross-sectional and longitudinal levels. At a cross-sectional level, multidimensional IRT models may allow deriving not only one, but two or more scales of deprivation (such as monetary and non-monetary). Different types of data and different models can be used (for example, polytomous IRT models describe and summarize other types of data than the dichotomous ones). The extension of the IRT at a longitudinal level could also be done in several directions and it may give insights into the process of change over time.

References:

[1] Baker, F.B., *Item Response Theory: Parameter estimation techniques*, New York: Marcel Dekker, 1992.

[2] Cappellari, L. and Jenkins, S.P., Multivariate probit regression using simulated maximum likelihood, *Stata Journal*, vol3, no.3, 2003, pp.278-294.

[3] Cappellari, L. and Jenkins, S.P., Summarizing Multiple Deprivation Indicators, in Jenkins S. and Micklewright J. (eds.), *Inequality and Poverty: Re-examined*, Oxford: Oxford University Press, 2007.

[4] Chiang, Y. and Wang, T., Applying Fuzzy Theory to the Management Competency Assessment for Middle Managers, *WSEAS Transactions on Business and Economics*, no.2, vol.4, 2007, pp.25-33.

[5] Dickes, P., Fusco, A., Rasch model and multidimensional poverty measurement, *IRISS Working Paper*, 2006.

[6] Dickes P., *Modèle de Rasch pour items dichotomiques : Théorie, Technique et application à la mesure de la pauvreté*, Université de Nancy II, 1983.

[7] Gailly, B., Hausman, P., Désavantagés relatifs a une Mesure Objective de la Pauvreté in SARPELLON G. (ed.), *Understanding Poverty*, Milan:Franco Angeli, 1984.

[8] Hardouin, J.B., Manual for the SAS macro-programs LoevH and MSP and the STATA modules LoevH and MSP, 2005.

[9] Hesketh, R., Skrondal, A. And Pickles, A., GLAMM Manual, UC Berkeley Division of Biostatistics, *Working Paper Series*, Working Paper no.160, 2004.

[10] Kakwani, N., Silber, J. *Quantitative approaches to multidimensional poverty measurement*, Palgrave Macmillan, 2007.

[11] Lord, F.M. *Applications of Item Response Theory to practical testing problems*, Hillsdale, NJ: Lawrence Erlbaum, 1960.

[12] Paraguas, F. and Kamil, A., Spatial Econometrics Modelling of Poverty, *WSEAS Transactions on Mathematics*, no.4, vol.4, 2005, pp.368-374.

[13] Rasch, G., An individualistic approach to item analysis, in P. Lazarsfeld and N. Henry (eds.), *Reading in mathematical social science*, MIT Press: Cambridge, 1966.

[14] Raileanu Szeles, M., The patterns and causes of social exclusion in Luxembourg, *IRISS Working Paper Series* 2007-09, IRISS at CEPS/INSTEAD.

[15] Safiih Muhamad, L., Basah Kamil, A.A. and Abu Osman M.T., Fuzzy Semi-parametric Sample Selection Model Case Study for Participation of Married Women,

WSEAS Transactions on Mathematics, no.3, vol.7, 2008, pp.112-117.

[16] Stout, W.F., A nonparametric approach for assessing latent trait unidimensionality, *Psychometrika*, no. 52, 1987, pp. 589–617.

[17] Szeles, M., Multidimensional Poverty Comparisons within Europe. Evidence from the European Community Household Panel, *IRISS Working Paper Series* 2004-05, IRISS at CEPS/INSTEAD.

[18] Wilson, M. And Boeck, P. (eds.), *Explanatory item response models: A generalized liniar and nonliniar approach*, Springer New York, 2004.

[19] Zheng, X, Rabe-Hesketh, S., Estimating parameters of dichotomous and ordinal item response models with gllamm, *The Stata Journal* 7, no.3, pp.313-333, 2007.