EKONOMETRIA ECONOMETRICS 2(52) • 2016

ISSN 1507-3866
e-ISSN 2449-9994

Justyna Brzezińska
University of Economics in Katowice
e-mail: justyna.brzezinska@ue.katowice.pl

POLYTOMOUS ITEM RESPONSE THEORY MODELS USING R

POLITOMICZNE MODELE TEORII ODPOWIEDZI NA POZYCJE TESTOWE W PROGRAMIE R

DOI: 10.15611/ekt.2016.2.04
JEL Classification: C1, C83.

Summary: Item response theory (IRT) is widely used in educational and psychological research to model how participants respond to test items in isolation and in bundles. Item response theory has replaced classical measurement theory as a framework for test development, scale constructions, scree reporting and test evaluation. The most popular of the item response models for multiple choice tests are the one-parameter (i. e. the Rasch model) and three-parameter models. This is the general framework for specifying the functional relationship between a respondent's underlying latent trait level, commonly known as ability in educational testing, or the factor score in the factor analysis tradition and an item level stimulus. In this paper, arguments are offered for continuing research and applying multidimensional IRT models. The position is also taken that multi-parameter IRT models have potentially important roles to play in the advancement of measurement theory about which models to use should depend on model fit to the test data. All calculations are conducted in R available from CRAN which is a widely-used and well-known environment for statistical computing and graphics.

Keywords: IRT models, multidimensional IRT models, measurement theory, R software.

Streszczenie: Teoria odpowiedzi na pozycje testowe (*Item Response Theory*) jest metodą statystyki wielowymiarowej, którą najczęściej wykorzystuje się w badaniach edukacyjnych oraz psychologicznych. Metoda ta pozwala na modelowanie cech ukrytych o charakterze ciągłym na podstawie dyskretnych wskaźników. Najczęściej wskaźnikami są odpowiedzi udzielone na zadania testowe oraz odpowiedzi na pytania kwestionariuszowe, rzadziej zaobserwowane cechy respondentów. Modele IRT wiążą cechę ukrytą ze wskaźnikami dzięki zastosowaniu parametryzacji, która określa właściwości wskaźników i rozkład cech respondentów. Jednym z najpopularniejszych modeli wielokrotnego wyboru jest model jednoparametryczny (model Rascha) oraz model trójparametryczny. W artykule tym zaprezentowano argumenty przemawiające za kontynuacją badań oraz zastosowaniem wielowymiarowych modeli w praktyce. Przedstawiono także, jak istotną rolę w teorii pomiaru odgrywają wielowymiarowe modele IRT. Obliczenia przedstawione w niniejszym artykule przeprowadzono w programie R.

Słowa kluczowe: modele teorii odpowiedzi na pozycje (IRT), wielowymiarowe modele IRT, teoria pomiaru, program R.

1. Introduction

Item response theory models (IRT) are increasingly becoming established in social science research, particularly in the analysis of performance or attitudinal data in psychology, education, medicine, marketing and other fields where testing is relevant.

Item response theory (IRT) models show the relationship between the ability or trait (symbolized by θ) measured by the instrument and an item response. The item response may be dichotomous (two categories), such as right or wrong, yes or no, agree or disagree. Or, it may be polytomous (more than two categories), such as a rating from a judge or scorer or a Likert-type response scale on a survey. The construct measured by the items may be an academic proficiency or aptitude, or it may be an attitude or belief.

The first generation of item response models developed in the 1940s and 1950s were intended to be applied to unidimensional test items that were dichotomously scored. Item response functions incorporated one, two, or three parameters and were one of two mathematical forms, normal-ogive or logistic. Lord [1952], introduced the two-parameter normal-ogive model for analyzing multiple-choice test data. Applications of his model were hampered by the complexities of model parameter estimation and Lord's concern about the failure to handle the problem of guessing in his two-parameter model. A few years later in the late 1950s, Birnbaum in a series of reports described the more tractable two and three-parameter logistic models. This important work is most accessible today in Birnbaum [1968]. While important technical and application work was going on with the unidimensional normal and logistic IRT models, other developments started in the late 1960s and became the serious activity of researchers beginning in the early 1980s. Samejima [1969], for example, introduced the very important graded response model to analyze data from Likert attitude scales and polynomously scored performance tasks such as might arise in the scoring of writing samples. Her model and variations on it were the first of many models developed by her and other scholars to handle ordered polytomous data. The work was stimulated by both the desire to generate and investigate new and potentially interesting models, and by an increase in the presence of polytomous data in educational testing.

Item Response Theory (ITR) is an extension of Classical Test Theory (CTT) [Lord, Novick 1968; Birnbaum 1968; Baker 1985; Bock, Lieberman 1970; Bock 1972; Lord 1980; Wright 1992; 1997]. The mathematical foundation of IRT is a function that relates the probability of a person responding to an item in a specific manner to the standing of that person on the trait that the item is measuring. It means that the function describes, in probabilistic terms, how a person with a higher standing on a trait is likely to provide a response in a different category to a person with a low standing on the trait. This mathematical function has a pre-specified form (usually a logistic ogive) and is referred to as an item response function (IRF).

The main advantage of IRT models is the fact that the item location parameter (b) and the person trait level (θ) are indexed on the same metric. Therefore, when a person`s trait level is higher than the item location on the trait continuum, that person is more likely than not to provide a trait-indicating (positive or true) response. The converse is true when a person`s trait level is below the item location. IRT models are well suited to cope with dichotomous and polytomous responses, where the response categories may be unordered as well as ordered. The incorporation of linear structures allows for modeling the effects of covariates and enables the analysis of repeated categorical measurements.

Polytomous items have become omnipresent in educational and psychological testing. Polytomous IRT models can be used for any test question where there are several response categories available. There are several types of polytomous IRT models: nominal response model, partial credit model, generalized partial credit model, rating scale model and graded response model. In this paper we present the most popular and best known IRT polytomous models. We also elaborate on the potential of the ltm package in R software (www.r-project.org). The ltm (latent trait models under IRT) package allows for the analysis of multivariate dichotomous and polytomous data using latent trait models under the Item Response Theory approach. It includes the Rasch, the Two-Parameter Logistic, Birnbaum's Three-Parameter, the Graded Response, and the Generalized Partial Credit Models.

In this paper we present the commonly known polytomous item response theory models. We show how these models originated and were developed, as well as how they have inspired the applied researchers and measurement practitioners. Additionally, we elaborate on the application of polytomous IRT models in R software.

2. Polytomous item response theory models

Items that are scored in two categories (e.g. right/wrong) are referred to as dichotomously scored items. Items scored in multiple-ordered categories are referred to as polytomously scored items. For the dichotomously scored items, the probability of a correct response for an examinee can be described by one of the logistic IRT models, most typically the three-parameter logistic (3PL) model IRT model if the items are multiple choice. For the polytomous scored items, the probability of an examinee reaching a specific score can be described by one of the polytomous IRT models, among which are the partial credit model [Masters 1982], and its generalized partial credit model – the graded response model [Samejima 1969; 1972].

Polytomous models are extensively used in applied psychological measurement and strongly related to the increase of statistical information when compared to dichotomous items. In some settings those models may help in reducing test length such time, costs, respondents' motivation, etc.

Polytomous IRT models are for items in which the categories are ordered, they cannot be used to determine the empirical ordering of the categories post hoc. They are appropriate for items or products (presentations, portfolios, essays, etc.) scored using a scoring rubric. They are also appropriate for Likert-type items, items with an ordered response scale such as: strongly disagree, disagree, neutral, agree, strongly agree. In these models a function analogous to an ICC can be plotted for each category. Unfortunately, the term used to label these curves is not universal, so the reader must infer it from the context in which the function is plotted.

2.1. Partial credit model

Masters [1982], introduced the partial credit model (PCM) as an IRT model for at least two polytomous items with ordered categories. The partial credit model can be considered as an extension of 1 parameter model and it has all the standard Rasch model features such as separability of person and item parameters. The partial credit model can be given if the respondents answered correctly to the first but not all the steps. In this type of model a varying number of categories across items is possible. By incorporating a location parameter (*b*) for each category boundary (*g*) and each item (*i*), we obtain a flexible model where categories can vary in number and structure across items within a test [Masters 1982]. The PCM can be described as the probability of responding in a specific item category (P_{i_g}):

$$P_{i_g}(\theta) = \frac{\exp\left[\sum_{g=0}^{l}\left(\theta - b_{i_g}\right)\right]}{\sum_{h=0}^{m}\left[\exp\sum_{g=0}^{h}\left(\theta - b_{i_g}\right)\right]}, \tag{1}$$

where: θ is latent variable (the person trait level), b_{i_g} is the location parameter of the category boundary function for category g of item i ($l = 0,...,g$, $h = 0,...,g,...,m$).

Each adjacent pair of response categories is modelled by a simple logistic ogive with the same slope to produce category boundaries that differ only in location. Because the PCM allows for a relatively small number of estimates per set of items, sample sizes as small as 300 return a stable item parameter and trait estimation [De Ayala 2009].

2.2. Graded response model

The graded response model (GRM) is an extension of Thurstone's [1928], method of successive intervals to the analysis of graded responses on educational tests. Samejima [1969], described a graded response IRT model as one in which an item

has m_j ordered response categories. The examinee is permitted to select only one of the categories.

The mathematical function for a graded response model looks very like the function for 2PL (2 parameters model) IRT model. The difference is that there are multiple *b*-parameters, one for each category except the first [Samejima 1969]:

$$P_{i_g}^{*}(\theta) = \frac{\exp\left[a_i \left(\theta - b_{i_g} \right) \right]}{1 + \exp\left[a_i \left(\theta - b_{i_g} \right) \right]}, \tag{2}$$

where: $P_{i_g}^{*}(\theta)$ is the probability of scoring in or above category *g* of item *i* (given θ and the item parameters), a_i is the item discrimination parameter, b_{i_g} is the boundary location parameter, or threshold for category *g* of item *i*.

The partial credit model describes the probability of reaching a score category by the difference of two probabilities, each of which can be expressed through the use of a dichotomous IRT model. A graded response model is suitable for a Likert-type rating. To fit this model within a measure, the item need not have the same number of response categories, no complications arise in item parameter estimation or the subsequent parameter interpretation as a result of a measure having items with different response formats.

2.3. Polytomous Rasch model

The Rasch model was proposed in the 1960s by the Danish statistician Georg Rasch. The basic Rasch model is used to separate the ability of test takers and the quality of the test [Rasch 1960; 1966;1977]. Rasch models have been routinely used with great success in order to build one-dimensional scales of unobserved quantities. The use of the Rasch model has been motivated mainly by its simplicity, compared to the Item Response Models (IRM). Its widespread use has also been motivated by the ability of the Rasch model to accommodate for sparse data matrices (with missing data) and to generate ratio measures of latent (unobserved) traits and abilities from ordinal observations. Rasch models were developed by Masters [1982], Bechtel [1985], Andrich [2004] and Christensen, Krelner and Mesbah [2013].

The Rasch model assigns one scale parameter to each person called ability, and one scale parameter to each item called item difficulty. Note that the raw score is a sufficient statistic for the Rasch estimates and that the person and item parameters have a common measurement unit which is called the "logit". The logit scale is linear, so differences of one logit, for example, have the same meaning at all the points on the scale. Since the items and persons appear on the same logit scale, it is convenient to observe that a person is expected to get the items below his/her ability level correct and those above his/her ability level incorrect. The probability of a correct response for the Rasch model may be given by the formula:

$$P_{ij}\left(\theta_j, b_i\right) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \tag{3}$$

where person j with ability θ_j has a probability P_{ij} to respond correctly to item i with difficulty b_i ($-\infty < b_i < \infty$). It is assumed that item i is dichotomously scored (0 for incorrect response, 1 for correct response). This is often called the simple Rasch model because it only models correct/incorrect responses, i.e. it does not model partially correct responses.

The use of the Rasch model has been motivated mainly by its simplicity compared to the Item Response Models (IRM). Its widespread use has also been motivated by the ability of the Rasch model to accommodate for sparse data matrices (with missing data) and to generate ratio measures of latent (unobserved) traits and abilities from ordinal observations.

Not all sets of data, however, can meet the requirements of the Rasch model. One of the fundamental assumptions of the Rasch model is that of Local Independence, which states that the observed responses of a person are independent of each other given an individual's position on the latent variable (i.e. on the logit scale). In some cases this assumption may not hold in practical settings, and the use of multilevel Rasch models may be a useful alternative.

3. Application in R

Polytomous Item Response Theory Models provide a unified, comprehensive introduction to the range of polytomous models available within item response theory (IRT). Analysis of Polytomous Item Response is available in the `ltm` package with the use of `gpcm` function in R (www.r-project.org) [Rizopoulos 2006]. The ltm provides a flexible framework for Item Response Theory analyses for dichotomous and polytomous data under the Marginal Maximum Likelihood approach.

In this paper we fit the Generalized Partial Credit model for ordinal polytomous data, under the Item Response Theory approach. We use the `Science` dataset that comes from the Consumer Protection and Perceptions of Science and Technology section of the 1992 Euro-Barometer Survey [Karlheinz, Melich 1992], based on a sample from Great Britain. All of the items below were measured on a four-group scale with response categories "strongly disagree", "disagree to some extent", "agree to some extent" and "strongly agree":

- Comfort: science and technology are making our lives healthier, easier and more comfortable.
- Environment: scientific and technological research cannot play an important role in protecting the environment and repairing it.
- Work: the application of science and new technology will make work more interesting.

- Future: thanks to science and technology, there will be more opportunities for the future generations.
- Technology: new technology does not depend on basic scientific research.
- Industry: scientific and technological research do not play an important role in industrial development.
- Benefit: the benefits of science are greater than any harmful effect it may have.

Coefficients of the different types of Generalized Partial Credit Model for the Science data including parameters and information criteria (AIC, BIC) are presented in Table 1. Only four out of the seven categories (Comfort, Work, Future, Benefit) were selected for further analysis.

Table 1. Coefficients of the Generalized Partial Credit Models

GPCM				
Coefficient	Catgr.1	Catgr.2	Catgr.3	Dscrmn
Comfort	–3.277	–2.891	1.537	0.861
Work	–2.035	–1.033	2.059	0.840
Future	–2.083	–0.975	0.832	2.234
Benefit	–2.908	–1.109	1.631	0.721
The log-likelihood	–1612.683			
AIC	3257.365			
BIC	3320.906			
GPCM assuming equal discrimination parameters across items				
Coefficient	Catgr.1	Catgr.2	Catgr.3	Dscrmn
Comfort	–3.083	–2.592	1.387	1.001
Work	–1.894	–0.910	1.134	1.001
Future	2.644	–1.420	1.134	1.001
Benefit	–2.447	–0.899	1.357	1.001
The log-likelihood	–1619.274			
AIC	3264.548			
BIC	3316.175			
GPCM assuming equal discrimination parameters across items fixed at 1				
Coefficient	Catgr.1	Catgr.2	Catgr.3	Dscrmn
Comfort	–3.085	–2.595	1.388	1
Work	–1.895	–0.911	1.858	1
Future	–2.646	–1.421	1.135	1
Benefit	–2.448	–0.899	1.357	1
The log-likelihood	–1619.274			
AIC	3262.548			
BIC	3310.204			

Source: own calculations in R.

In the next step of the analysis we built another type of polytomous IRT model – the Graded Response Model. The analysis of this model is available in R with the use of the `grm` function in the `ltm` library. The coefficients of the Graded Response Model and information criteria (AIC, BIC) are presented in Table 2.

Table 2. Coefficients of the Graded Response Model

GRM				
Coefficient	Extrmt1	Extrmt2	Extrmt3	Dscrmn
Comfort	–4.672	–2.536	1.408	1.041
Work	–2.385	–0.735	1.849	1.226
Future	–2.281	–0.965	0.856	2.299
Benefit	–3.060	–0.906	1.543	1.094
The log-likelihood	–1608.871			
AIC	3249.742			
BIC	3313.282			
GRM assuming equal discrimination parameters across items				
Coefficient	Extrmt1	Extrmt2	Extrmt3	Dscrmn
Comfort	–3.910	–2.153	1.201	1.321
Work	–2.264	–0.707	1.756	1.321
Future	–3.079	–1.258	1.126	1.321
Benefit	–2.687	–0.800	1.367	1.321
The log-likelihood	–1613.899			
AIC	3253.798			
BIC	3305.425			

Source: own calculations in R.

The two focal polytomous models of interest to the current project are the GPCM and GRM for Science dataset. The GRM manifests itself as a proportional odds model in which for each item, all response categories are collapsed into two categories when estimating the IRFs. Therefore the two models do not indicate the same ordering among score categories and do not produce directly comparable parameters [Ostini, Nering 2005], although many have found that these common polytomous IRT models tend to produce very similar results.

The likelihood ratio test, Akaike's information criterion AIC [Akaike 1974] and Bayesian information criterion BIC [Schwarz 1978], were calculated to compare the fit of the two models. In this case, AIC and BIC for Graded Response Model (3249.742 and 3313.282, respectively) are lower than those of the Generalized Partial Credit Model.

4. Conclusions

Polytomous item response theory models provide a unified, comprehensive introduction to the range of polytomous models available within item response theory (IRT). It begins by outlining the primary structural distinction between the two major types of polytomous IRT models. This focuses on the two types of response probability that are unique to polytomous models and their associated response functions, which are modeled differently by the different types of IRT model. In this paper we present both conceptually and mathematically, the major specific polytomous IRT models. Relationships among the models are also investigated and the operation of measurement information is described for each major model. Practical examples of major models using real data are provided, as is a chapter on choosing the appropriate model.

There are some disadvantages of IRT models such as: the strict assumptions to be followed, they are more difficult to use then CTT, and they are more complex and difficult to understand.

The purpose of this paper was to provide a summary of some of the latest developments in polytomous item response theory (IRT), and to help realize that psychometric tools can now be used for theory testing in addition to the traditional role of improving construct measurement. In this study, the application of general polytomous item response theory models, as well as practical considerations have been presented. We presented the `gpcm` and `grm` package available in R that covers the most important features of the analysis for poltomous IRT models. The main functions of these packages have been presented and a comparison over the presented polytomous IRT models was discussed. Coefficients for several poltytomous IRT models were calculated, as well as the information criteria for testing the goodness of fit. For model selection we applied the likelihood ratio test, AIC and BIC to compare the fit of alternative models to these data. Other fit statistics may also be useful in investigating the local item dependence present in these data. In testing the goodness of fit we chose the model with the lowest information criterion indicating the best fitting model.

Bibliography

Akaike H., 1974, *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 19 (6), pp. 716–723.

Andrich D., 2004, *Controversy and the Rasch model: a characteristic of incompatible paradigms?,* Medical Care, 42 (1 Supplement), pp. I–7.

Baker F.B., 1985, *The basic of item response theory*, College Park, MD: ERIC Clearinghouse on Assessment and Evaluation.

Bechtel, G.G., 1985, *Generalizing the Rasch model for consumer rating scales*, Marketing Science, 4 (1), pp. 62–73.

Birnbaum A., 1968, *Some Latent Trait Models and Their Use in Inferring an Examinee`S Ability*, [in:] *Statistical Theories of Mental Test Scores*, eds. F.M. Lord, M.R. Novick, Addison-Wesley, Reading, pp. 395–479.

Bock R.D., 1972, *Estimating item parameters and latent ability when response are scored in two or more nominal categories*, Psychometrika, 37, pp. 29–51.

Bock R., Lieberman M., 1970, *Fitting a response model for n dichotomously scored items*, Psychometrika, 35, pp. 179–197.

Christensen K.B., Kreiner S., Mesbah M., 2013, *Rasch Models in Health*, London–Hoboken: ISTE––Wiley.

De Ayala R.J., 2009, *Theory and Practice of Item Response Theory*, Guilford Publications.

Karlheinz R. and Melich A., 1992, Euro-Barometer 38.1: *Consumer Protection and Perceptions of Science and Technology*, INRA (Europe), Brussels [computer file].

Lord F.M., 1952, *Theory of test scores*, Psychometric Monographs, no. 7.

Lord F.M., 1980, *Applications of Item Response Theory to Practical Testing Problems*, Hillsdale: Lawrence Erlbaum.

Lord F.M., Novick M.R., 1968, *Statistical theories of mental test scores* (with contributions by A. Birnbaum), Reading, MA: Addison-Wesley.

Masters G.N., 1982, *A Rasch model for partial credit scoring*, Psychometrika, 47, pp. 149–174.

Masters G.N., Wright B. D. (1984), The essential process in a family of measurement models, *Psychometrika,* 49, pp. 529–544.

Ostini, R., Nering, M. L. (2005), *Polytomous item response theory models*. Thousand Oaks: Sage.

Rasch G., 1960, *Probabilistic Models for some Intelligence and Attainment Tests*, Danish Institute for Education Research, Copenhagen.

Rasch G., 1966, *An Individualistic Approach to Item Analysis*, [in:] P.F. Lazarsfeld, N.W. Henry (eds.), *Readings in Mathematical Social Sciences*, Cambridge: MIT Press, pp. 89–107.

Rasch G., 1977, *On specific objectivity: An attempt at formalising the request for generality and validity of scientific statements*, Danish Yearbook of Philosophy, 14, pp. 58-94.

Rizopoulos D., 2006, *ltm: An R package for latent variable modelling and item response theory analyses*, Journal of Statistical Software, 17(5), pp. 1–25, http://www.jstatsoft.org/v17/i05/.

Samejima F.,1969, *Estimation of latent ability using a response pattern of graded scores*, Psychometric Monograph, no. 17.

Samejima F., 1972, *A general model for free-response data*, Psychometrika Monograph, no. 18.

Schwarz G., 1978, *Estimating the dimension of a model*, Annals of Statistics, 6, pp. 461–464.

Thurstone L.L., 1928, *Attitudes can be measured*, American Journal of Sociology, 33, pp. 529–54.

Wright B.D., 1992, *IRT in the 1990s: which models work best? 3PL or Rasch*?, Rasch Measurement Transactions, 6 (1), pp. 196–200.

Wright B.D.,1997, *A history of social science measurement*, Educational Measurement: Issues and Practice, 16(4), pp. 33–45.