DE GRUYTER OPEN

ECONOMETRICS. ADVANCES IN APPLIED DATA ANALYSIS
Year 2018, Vol. 22, No. 1
ISSN 1507-3866
e-ISSN 2449-9994

ITEM RESPONSE THEORY MODELS IN THE MEASUREMENT THEORY WITH THE USE OF LTM PACKAGE IN R

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

DOI: 10.15611/eada.2018.1.01

Abstract: Item Response Theory (IRT) is an extension of the Classical Test Theory (CCT) and focuses on how specific test items function in assessing a construct. They are widely known in psychology, medicine, and marketing, as well as in social sciences. An item response model specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. Within the broad framework of item response theory, many models can be operationalized because of the large number of choices available for the mathematical form of the item characteristic curves. In this paper we introduce several types of IRT models such as: the Rasch, and the Birnbaum model. We present the main assumptions for IRT analysis, estimation method, properties, and model selection methods. In this paper we present the application of IRT analysis for binary data with the use of the `ltm` package in R.

Keywords: Item Response Theory (IRT), measurement theory, latent class analysis, R software.

1. Introduction

Item Response Theory shows the relationships between the ability or trait (θ) measured by the instrument and an item response. The item may be dichotomous when we deal with two categories, or it may be polytomous for more than two categories. Since traits are not directly measurable, they are referred to as latent traits or abilities. An item response model specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. Within the broad framework of item response theory, many models can be operationalized because of the large number of choices available for the mathematical form of the item characteristic curves. But, whereas item response theory cannot be shown to be correct or incorrect, the appropriateness of particular models with any set of test data can be established by conducting

a suitable goodness -of-fit investigation. The relationship between the observable and the unobservable quantities is described by a mathematical function. For this reason, item response models are mathematical models which are based on specific assumptions about the test data. Different models, or item response models as they are called, are formed through specifying the assumptions one is willing to make about the test data set under investigation.

There are three primary advantages of item response models:

1. Assuming the existence of a large pool of items all measuring the same trait, the estimate of an examinee's ability is independent of the particular sample of test items that are administered to the examinee.

2. Assuming the existence of a large population of examinees, the descriptors of a test item (for example, item difficulty and discrimination indices) are independent of the particular sample of examinees drawn for the purpose of calibrating the item.

3. A statistic indicating the precision with which each examinee's ability is estimated is provided. This statistic is free to vary from one examinee to another. Needless to say, the extent to which the three advantages are gained in an application of an item response model depends on the closeness of the "fit" between a set of test data and the model. If the fit is poor, these three desirable features either will not be obtained or will be obtained in a low degree. An additional desirable feature is that the concept of parallel forms reliability is replaced by the concept of statistical estimation and associated standard errors.

The goal of item response theory is to provide both invariant item statistics and ability estimates. These features will be obtained when there is a reasonable fit between the chosen model and the data set. Through the estimation process, items and persons are placed on the ability scale in such a way that there is as close a relationship as possible between the expected examinee probability parameters and the actual probabilities of performance for examinees positioned at each ability level. Item parameter estimates and examinee ability estimates are revised continually until the maximum agreement possible is obtained between predictions based on the ability and item parameter estimates and the actual test data. Research in the field of item response theory models include works by Andrich [1978], Rasch [1960], Birnbaum [1968], Lord and Novick [1968], Samejima [1969] and Bock [1972]. The latest and modern approach present Reckase [2009], DeBoeck [2004], Bates [2008], DeMars [2010] and Chalmers [2012].

2. Item Response Theory

2.1. Background to Item Response Theory models

The common models and procedures for constructing tests and interpreting test scores have served measurement specialists and other test users well for a long time. These models, such as the classical test model, are based on weak assumptions, that

is, the assumptions can be met easily by most test data sets, and, therefore, the models can and have been applied to a wide variety of test development and test score analysis problems. Today, there are countless numbers of achievement, aptitude, and personality tests that have been constructed with these models and procedures. Well-known classical test model statistics, such as the standard error of measurement, the Spearman-Brown formula, and the Kuder-Richardson formula-20, are just a few of the many important statistics that are a part of the classical test model and related techniques [Gulliksen 1950; Lord, Novick 1968].

Item response theory (IRT) models show the relationship between the ability or trait (symbolized as y) 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 IRT score is often called an ability, trait or proficiency. The IRT scoring takes into account the item difficulty and discrimination. Items that are more discriminating, or more reliable, are weighted more heavily, so IRT scores can be more reliable than number-correct scores. If different examinees take different tests, the IRT scores adjust for the difference in difficulty. Item response theory also provides an index of the precision of the test score − the standard error of measurement − for each examinee.

Any theory of item responses supposes that, in testing situations, examinee performance on a test can be predicted (or explained) by defining examinee characteristics, referred to as traits, or abilities; estimating scores for examinees on these traits (called "ability scores"); and using the scores to predict or explain item and test performance [Lord, Novick 1968]. Since traits are not directly measurable, they are referred to as latent traits or abilities. An item response model specifies a relationship between the observable examinee test performance and the unobservable traits or abilities assumed to underlie performance on the test. Within the broad framework of item response theory, many models can be operationalized because of the large number of choices available for the mathematical form of the item characteristic curves. But, whereas item response theory cannot be shown to be correct or incorrect, the appropriateness of particular models with any set of test data can be established by conducting a suitable goodness-of-fit investigation.

The characteristics of an item response model are as follows [Hambleton, Swaminathan 1985]:

- it is a model which supposes that examinee performance on a test can be predicted (or explained) in terms of one or more characteristics referred to as traits,
- an item response model specifies a relationship between the observable examinee item performance and the traits or abilities assumed to underlie performance on the test,

- a successful item response model provides a means of estimating scores for examinees on the traits,
- the traits must be estimated (or inferred) from observable examinee performance on a set of test items (it is for this reason that there is the reference to latent traits or abilities).

The main applications of IRT can be found in educational testing in which analysis are interested in measuring examinees' ability using a test that consists of several items (i.e. questions). Several models and estimation procedures have been proposed that deal with various aspects of educational testing.

2.2. Item Response Theory models assumptions

There are three assumptions of Item Response Theory (IRT): unidimensionality, local independence, and correct model specification. A test that is unidimensional consists of items that tap into only one dimension. Whenever only a single score is reported for a test, there is an implicit assumption that the items share a common primary construct. Multidimensional IRT models exist, but they are not addressed here. Unidimensionality means that the model has a single y for each examinee, and any other factors affecting the item response are treated as random error or nuisance dimensions unique to that item and not shared by other items. Violating this assumption may lead to misestimation of parameters or standard errors.

Many methods have been proposed for testing unidimensionality. Hattie [1984] compared 87 methods. More recently, Tate [2003] compared nine of the most commonly used methods. Three common methods are the most popular and known nowadays: analysis of the eigenvalues of the inter-item correlation matrix, Stout's test of essential unidimensionality, and indices based on the residuals from a unidimensional solution. Another assumption of IRT is local independence. If the item responses are not locally independent under a unidimensional model, another dimension must be causing the dependence. With tests of local independence, however, the focus is on dependencies among pairs of items. These dependencies might not emerge as separate dimensions, unless they influenced a larger group of items, and thus might not be detectable by tests of unidimensionality. Consequently, separate procedures have been developed to detect local dependencies. If items are locally independent, they will be uncorrelated after conditioning on θ. Again, note that the items can (and should) be correlated in the sample as a whole. It is only after controlling for θ that we assume they are uncorrelated. Yen [1984] proposed a simple test, Q_3 , to check pairs of items for local dependence. Additional indices have been proposed for testing local independence [Chen, Thissen 1997], however Q_3 is one of the more commonly used.

The fit between the model and the data can be assessed to check for model misspecifications. For example, if a 1PL model is used and the data follow a model with varying slopes or a non-zero lower asymptote, then many of the items will not

fit the 1PL model. If the function is not monotonically increasing, none of the common models will fit. Typically, IRT practitioners focus on the fit of individual items, not the overall fit of the model across all items, which will only be addressed briefly at the end of this section. For item fit, the concept of a residual, or the difference between an observed and model-predicted (expected) proportion is key. This residual is conditional on θ, meaning it is calculated for groups of examinees with approximately the same θ. We can use an Item Characteristic Curve (ICC) which represents the model expectation. We can also use the Pearson χ^2, log-likelihood G^2, or AIC and BIC information criterion.

2.3. IRT models for dichotomous items

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 assigned to measure more than one trait and these models are referred to as unidimensional IRT. In an educational testing situation in which \$n\$ individuals answer I questions for items. For $j = 1,...,n$ and $i = 1,...,I$, let Y_{ij} be random variables associated with the response of individual j to item i. These respondents may be binary (correct or incorrect answer) or may be discrete with a number of categories. Let Ω_y denote the set of possible values of the Y_{ij}, assumed to be identical for each item in the test. Let θ_j denote the latent trait of ability for individual j, and let η_i denote a set of parameters that will be used to model item (question) characteristics. Different IRT models arise from different sets of possible responses Ω_y and different functional forms assumed to describe the probabilities with the Y_{ij} assume those values, namely:

$$P(Y_{ij} = y \mid \theta_j, \eta_i) = f(y \mid \theta_j, \eta_i);\ y \in \Omega_Y . \tag{1}$$

The item parameters η_i may include four distinct types of parameters: a discrimination parameter a_i, a difficulty parameter b_i, guessing parameter c_i, and carelessness parameter d_i. Depending on the number of parameters included in the model equation, there are several types of IRT models distinguished.

One parameter (1PL) IRT model is the simplest IRT model called the Rasch model [Rasch 1960]:

$$P(Y_{ij} = 1 \mid \theta_j, b_i) = \frac{\exp(\theta_j - b_i)}{1 + \exp(\theta_j - b_i)}, \tag{2}$$

where b_i $(-\infty < b_i < \infty)$ is difficulty (location, threshold) parameter and this value tells us how easy or how difficult an item is. In the Rasch model formulation (2) θ_j and b_i can take all real values and measure ability and difficulty respectively. The sign of expression $\theta_j - b_i$ in any particular instance indicates the probable outcome of the person-item interaction. If $\theta_j - b_i > 0$ then the most probable outcome is a correct response. If $\theta_j - b_i < 0$ then the most likely outcome is incorrect response.

Specific objectivity is the requirement that the measures produced by a measurement model be sample-free for the agents (test items) and test-free for the objects (people). Sample-free measurement means item difficulty estimates are as independent as is statistically possible of whichever persons, and whatever distribution of personal abilities, happen to be included in the sample. Test-free measurement means that person ability estimates are as independent as is statistically possible of whichever items, and whatever distribution of item difficulties, happen to be included in the test. In particular, the familiar statistical assumption of a normal (or any known) distribution of model parameters is not required.

This also implies that Rasch point-estimates are invariant when the data fit the Rasch model. The argument for invariance may be stated rather loosely as follows. Irrelevancies in the data should not make a fundamental difference in the results obtained from the analysis of the data. For Rasch measurement, irrelevancies include the person and item distributions.

Another IRT model is the 2-parameter (2PL) Birnbaum model defined as follows [Birnbaum 1957; Birnbaum 1958; Birnbaum 1968]:

$$P(Y_{ij}=1\left|\theta_j, a_i, b_i\right.) = \frac{\exp a_i\left(\theta_j - b_i\right)}{1+\exp a_i\left(\theta_j - b_i\right)}, \tag{3}$$

where a_i $(-\infty < a_i < \infty)$ is discrimination (slope) parameter. This parameter is related to how rapidly the probability changes with the changes in ability θ_j.

Birnbaum substituted the two-parameter logistic cumulative distribution function for the two-parameter normal ogive function as the form of the item characteristic curve. Logistic curves have the important advantage of being more convenient to work with than normal ogive curves. The 2PL model might be more appropriate for dichotomous attitudinal items; it may also be useful for multiple choice items with very effective distractors, where low-ability examinees would tend to think a particular distractors was right rather than to guess randomly.

Another 3-parameter (3PL) Birnbaum IRT model:

$$P(Y_{ij}=1\left|\theta_j, a_i, b_i, c_i\right.) = c_i + \left(1-c_i\right)\frac{\exp a_i\left(\theta_j - b_i\right)}{1+\exp a_i\left(\theta_j - b_i\right)}, \tag{4}$$

where c_i $(0 \le c_i \le 1)$ is the guessing parameter. In practice values of c_i above 0.35 are not considered acceptable. This value tells us how likely the examinees are to obtain the correct answer by guessing. A three-parameter (3Pl) IRT model uses three parameters a_i, b_i and c_i.

The parameter c_i is the lower asymptote of the item characteristic curve and represents the probability of examinees with low ability of correctly answering an item. The parameter is included in the model to account for item response data from low-ability examinees, where, among other things, guessing is a factor in test

performance. It is now common to refer to the parameter c_i as the pseudo-chance level parameter in the model. Typically, c_i assumes values that are smaller than the value that would result if examinees of low ability were to guess randomly to the item. As Lord [1974] noted, this phenomenon can probably be attributed to the ingenuity of item writers in developing "attractive" but incorrect choices. Low ability examinees are attracted to these incorrect answer choices. They would score higher by randomly guessing the correct answers. For this reason, avoidance of the label "guessing parameter" to describe the parameter c_i seems desirable. The 3PL model is the most common choice for multiple-choice items because it seems reasonable to assume that low-ability examinees have some non-zero probability of choosing the correct answer.

High-ability examinees do not always answer test items correctly. Sometimes these examinees may be a little careless, other times they may have information beyond that assumed by the test item writer; so they may choose answers that are not "keyed" as correct. To handle this problem, McDonald [1967] and more recently Barton and Lord [1981] have thus described a four-parameter (4PL) logistic model:

$$P(Y_{ij}=1\Big|\theta_j, a_i, b_i, c_i, d_i) = c_i + (d_i - c_i)\frac{\exp a_i(\theta_j - b_i)}{1+\exp a_i(\theta_j - b_i)}, \tag{5}$$

with carelessness parameter $d_i\,(0 \le c_i \le d_i \le 1)$.

This model differs from the three-parameter model in that d_i assumes a value slightly below 1. This model may be of theoretical interest only because Barton and Lord [1981] were unable to find any practical gains that accrued from the model's use.

The main application of IRT models can be found in education testing in which we measure examinees` ability using a test that consists of several items. Item response theory has a number of potential advantages over classical test theory in assessing self-reported health outcomes. IRT models yield invariant item and latent trait estimates (within a linear transformation), standard errors conditional on trait level, and trait estimates anchored to item content. IRT also facilitates the evaluation of differential item functioning, inclusion of items with different response formats in the same scale, and the assessment of person fit and is ideally suited for implementing computer adaptive testing. Finally, IRT methods can be helpful in developing better health outcome measures and in assessing change over time. These issues are reviewed, along with a discussion of some of the methodological and practical challenges in applying IRT methods.

2.4. Parameters estimates

In this subsection we provide a brief explanation of how the parameters in an IRT model for dichotomous items are estimated. Two types of estimation need to be performed: estimating the item parameters, and predicting the individual latent score θ.

Estimation of item parameters is done by using the marginal maximum likelihood (MML), or joint maximum likelihood (JML) estimation procedure. In marginal maximum likelihood, the estimation begins with the assumption of the distribution for θ, usually as a standard normal distribution. After the item parameters are estimated, the person parameters are estimated using maximum likelihood, or Bayesian-like approach via the mode or the mean of the posterior distribution, Modal A Posteriori (MAP) or Expected A Posteriori (EAP).

Consider an education testing situation in which n individuals answer I questions or items. Let Y_{ij} be a random variable associated with the binary response of individual j $(j = 1,...,n)$ to item i $(i = 1,...,I)$. Let Ω_y denote the set of possible values of the Y_{ij} for person j, with Ω_y are assumed to be identical for each item in the test. Let θ_j denote the latent trait of ability for individual j, and let η_j denote a set of parameters that will be used to model item (question) characteristics. Different IRT models arise from different sets of possible responses Ω_y and different functional forms assumed to describe the probabilities with which the Y_{ij} assume those values, namely:

$$P(Y_{ij} = y_{ij} \left| \theta_j, \eta_i \right.) = f\left(y_{ij} \left| \theta_j, \eta_i \right. \right). \tag{6}$$

Letting $\mathbf{y}_j = \left(y_{1j}, y_{2j}, ..., y_{ij}, ... y_{Ij} \right)$ be a vector of I observed binary responses from the j-th subject having an ability θ_j, the likelihood equation for person j is defined as:

$$L(\theta_j, \boldsymbol{\eta} \left| y_{ij} \right.) = P(y_{ij} \left| \theta, \boldsymbol{\eta} \right.) = \prod_{i=1}^{I} P(y_{ij} \left| \theta, \boldsymbol{\eta} \right.)^{y_{ij}} \left[1 - P\left(y_{ij} \left| \theta, \boldsymbol{\eta} \right. \right) \right]^{1-y_{ij}}, \tag{7}$$

where $\boldsymbol{\eta}$ is the vector of item parameters.

The likelihood function for all persons is defined as:

$$L(\boldsymbol{\theta}_j, \boldsymbol{\eta} \left| y_{ij} \right.) = \prod_{j=1}^{n} \prod_{i=1}^{I} P(y_{ij} \left| \theta, \boldsymbol{\eta} \right.)^{y_{ij}} \left[1 - P\left(y_{ij} \left| \theta, \boldsymbol{\eta} \right. \right) \right]^{1-y_{ij}}. \tag{8}$$

The full log-likelihood for n persons is defined as:

$$l(\boldsymbol{\theta}_j, \boldsymbol{\eta} \left| y_{ij} \right.) = \sum_{j=1}^{n} \sum_{i=1}^{I} y_{ij} \log\left(P\left(y_{ij} \left| \theta, \boldsymbol{\eta} \right. \right) \right) + \left(1 - y_{ij} \right) \log\left(1 - P\left(y_{ij} \left| \theta, \boldsymbol{\eta} \right. \right) \right). \tag{9}$$

Marginal maximum likelihood method utilizes the marginal distribution of the full log-likelihood of the item parameter obtained by integrating out θ.

Estimating the item parameter

The log-likelihood of the marginal distribution of the item parameters (9) given the form of the distribution $g(\theta_j)$ for the independent and identically distributed latent traits, is defined as:

$$l(\boldsymbol{\eta}) = \sum_{j=1}^{n} \sum_{i=1}^{I} \log \int \left[y_{ij} \log\left(P\left(y_{ij} \left| \theta, \boldsymbol{\eta} \right. \right) \right) + \left(1 - y_{ij} \right) \log\left(1 - P\left(y_{ij} \left| \theta, \boldsymbol{\eta} \right. \right) \right) \right] g\left(\theta_j \right) d\left(\theta_j \right). \tag{10}$$

The integral can be approximated using the Gauss-Hermit quadrature rule. The goal is to find the values of the components in $\boldsymbol{\eta}$ that maximize the integrated likelihood with respect to θ_j. Due to the local independence assumption it is possible to work with one item at a time. Within each item, the parameters are not independent so the maximums must be found simultaneously. We can use prior distributions for the item parameters and apply the Bayesian approach to estimate them. After the item parameters have been estimated, the distribution for θ can be updated. Then, the process of item estimation is repeated and updating the latent trait distribution until the components of $\boldsymbol{\eta}$ converge. The most updated distribution of θ is considered the posterior distribution, which can be used in the next step in the process of estimation the individual θ_j scores.

Estimating the individual latent scores θ_j s

Having the item parameters and the θ distribution estimated, the θ score for each subject is estimated using ML, Expected a Posteriori (EAP), or Modal a Posteriori (MAP) estimation method. Each individual (respondent) j has its own θ posteriori distribution, $g(\theta|\boldsymbol{\eta}, \mathbf{y}_j)$ which can be used in estimation of θ_j.

The Expected a Posteriori (EAP) estimation method uses the Gauss-Hermit quadrature rule to approximate the mean of the distribution,

$$\hat{\theta} = E\left[g\left(\theta|\boldsymbol{\eta}, \mathbf{y_j}\right)\right]. \tag{11}$$

The EAP estimation procedure estimates θ_j by using the mean of the distribution as the expected value. It is non-iterative method, where under the quadrature approximation approach, $g(\theta|\boldsymbol{\eta}, \mathbf{y}_j)$ may be approximated by finding the area under the curve of the function via a discrete distribution such as histogram.

In Modal a Posteriori (MAP), the mode is found by applying the Fisher Scoring Method defined as:

$$\hat{\theta}_{t+1} = \hat{\theta}_t - \left[\frac{\dfrac{\partial l\left(\theta_j|\boldsymbol{\eta}, y_{ij}\right)}{\partial \theta_j}}{\dfrac{\partial^2 l\left(\theta_j|\boldsymbol{\eta}, y_{ij}\right)}{\partial \theta_j^2}}\right]_t, \tag{12}$$

where $\dfrac{\dfrac{\partial l\left(\theta_j|\boldsymbol{\eta}, y_{ij}\right)}{\partial \theta_j}}{\dfrac{\partial^2 l\left(\theta_j|\boldsymbol{\eta}, y_{ij}\right)}{\partial \theta_j^2}}$ is the ratio of the first derivative of the log-likelihood function of θ_j and the Hessian matrix as a matrix of second derivatives of the log-likelihood function of θ_j. MAP is an iterative estimation method with $\hat{\theta}_t$ being updated until convergence is achieved.

3. Application in R

Over the years there has been an extensive growth in computer programs that can conduct item response theory, and within R there are at least several powerful packages: `eRm` [Mair, Hatzinger 2007], `ltm` [Rizopoulos 2006], `lme4` [Doran, Bates, Bliese 2007], `mirt` [Chalmers 2012], and `MiscPsycho` [Doran 2010]. Additional packages include `mokken` [Van der Ark 2010] to do non-metric IRT, `plink` [Weeks 2010] to link multiple groups together and the newest package `TAM` [Robitzsch, Kiefer, Wu 2017].

In this paper we use Abortion data-set from the `ltm` package in R [McGrath, Waterton 1989]. The data contain responses given by 410 individuals to four out of seven items concerning attitudes to abortion. A small number of individuals did not answer some of the questions and this data set contains only the complete cases.

In the initial step, we can use descriptive statistics for the Abortion dataset using the `descript` function: `descript(Abortion)`. The abortion data-set contains 4 items and 379 sample units; there are 0 missing values.

For the Abortion data we observe that item 3 seems to be the easiest one having the highest proportion of correct responses, while only item 1 indicates having a low degree of association. The frequencies of total scores are respectively: 103 for 0, 33 for 1, 37 for 2, 65 for 3, and 141 for 4. Table 1 presents the proportions for each level of response for the analyzed dataset on abortion.

Table 1. Proportions for each level of response

	0	1	Logit
Item 1	0.5620	0.4380	–0.2493
Item 2	0.4063	0.5937	0.3791
Item 3	0.3641	0.6359	0.5575
Item 4	0.3826	0.6174	0.4786

Source: own calculations in R.

We initially fit the Rasch model `Rasch` function using `fit<-rasch(Abortion)` function in R. Parameter estimates for the Rasch models, standard error and *z*-values are presented in Table 2.

Table 2. Coefficients for the Rasch model

	Value	Standard error	*z*-value
Difficulty Item 1	0.1636	0.0605	2.7040
Difficulty Item 2	–0.2366	0.0589	–4.0143
Difficulty Item 3	–0.3581	0.0617	–5.8069
Difficulty Item 4	–0.3039	0.0607	–5.0096
Discrimination	4.4571	0.3902	11.4217

Source: own calculations in R.

To inspect the main properties of the model, a `summary` can be called on. In the model summary the fitted log-likelihood and the information criteria AIC and BIC are reported: log.Lik = –708.5504, AIC = 1427.101, and BIC = 1446.789. To obtain details in R we can use the following command: `summary(fit)`.

The parameter estimates can be transformed to probability estimates using the `coef(fit, prob = TRUE, order = TRUE)` with detailed results presented below (Table 3). The column P($x = 1|z = 0$) denotes the probability of a positive response to the *i*-th item for the average individual. The order argument can be used to sort the items according to the difficulty estimates.

In order to check the fit of the model to the data, the `GoF.rasch` and `margins` functions can be used. The `GoF.rasch` function performs a parametric Bootstrap goodness-of-fit test using Pearson's χ^2 statistic. In particular, the null hypothesis states that the observed data have been generated under the Rasch model with parameter values the maximum likelihood estimates $\hat{\theta}$.

To test this hypothesis B samples are generated under the Rasch model using $\hat{\theta}$, and the Pearson's χ^2 statistic T_b ($b = 1,...,B$) is computed for each data-set; the *p*-value is then approximated by the number of times $T_b > T_{obs}$ plus one, divided by B +1, where T_{obs} denotes the value of the statistic in the original data-set. For the Abortion data this procedure yields the following results: $T_{obs} = 23.64$ for the sample 200, *p*-value = 0.01 suggests an acceptable fit of the model.

Table 3. Probability estimates

	Difficulty	Discrimination	$P(x = 1\| z = 0)$
Item 1	–0.3581	4.4571	0.8315
Item 2	–0.3039	4.4571	0.7948
Item 3	–0.2366	4.4571	0.7416
Item 4	0.1636	4.4571	0.3253

Source: own calculations in R.

Secondly, we test whether incorporating a guessing parameter to the unconstrained Rasch model improves the fit. This extension can be fitted using `tpm`, which has syntax very similar to `rasch` and allows one to fit either a Rasch model with a guessing parameter or the three-parameter IRT model. To fit the unconstrained Rasch model with a guessing parameter we obtain the following coefficients (Table 4).

Table 4. Coefficients for the Rasch model with guessing parameter

	Guessing	Difficulty	Discrimination
Item 1	0.000	0.149	4.997
Item 2	0.022	–0.204	4.997
Item 3	0.000	–0.349	4.997
Item 4	0.054	–0.233	4.997

Source: own calculations in R.

To compare two models we can use `ANOVA` function giving the result presented in likelihood ratio table (Table 5).

Table 5. Likelihood ratio table

	AIC	BIC	log.Lik	LRT	df	*p*-value
Rasch unconstrained	1427.10	1446.79	–708.55			
Rasch with guessing parameter	1434.02	1466.46	–708.01	1.08	4	0.898

Source: own calculations in R.

The definitions of AIC and BIC used by the `summary` and `anova` methods in `ltm` are such that "smaller is better". In this example on Abortion data we choose the Rasch unconstrained model as fitting better.

Adopting the unconstrained Rasch model, as well as the Rasch model including guessing parameter, we produce the Item Characteristic, and the Item Information Curves, by appropriate calls to the `plot` method for class `rasch`. Below, we present Item Information Curves and Item Characteristic Curves for unconstrained Rasch model (Figure 1) and for the unconstrained Rasch model with guessing parameter (Figure 2).

According to the Test Information Curve presented in Figure 1 we can see that the items about asked in the Abortion data mainly provide information for respondents with lower ability. In particular, the amount of test information for ability levels in the interval (−4,0) is around 50%, whereas the item that seems to distinguish between respondents with higher ability from the interval (0,4) levels is another 50%.

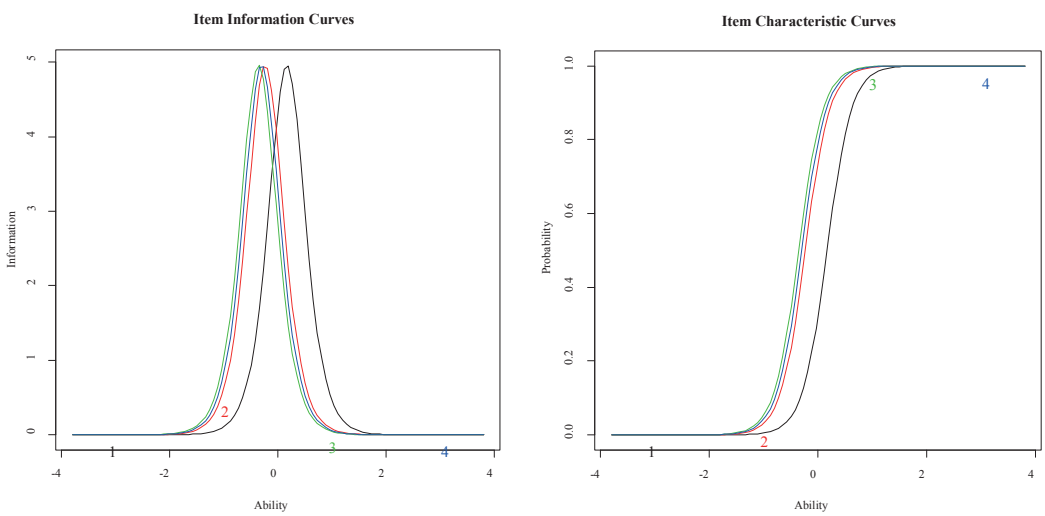

Figure 1. Item Information Curves (IIC) and Item Characteristic Curves (ICC) for unconstrained Rasch model

Source: own calculations in R.

Comparing plots for the unconstrained Rasch model and the model including guessing parameter we can see that there is no a big difference in ability.

In the last part of the analysis we can use `factor.scores` function for best fitting model that is the unconstrained Rasch model.

Table 6. Factor scores for observed response patterns

	Item 1	Item 2	Item 3	Item 4	Observed	Expected	$z1$	SE($z1$)
1	0	0	0	0	103	103.584	–0.903	0.458
2	0	0	0	1	13	11.713	–0.444	0.254
3	0	0	1	0	10	14.916	–0.444	0.254
4	0	0	1	1	21	14.079	–0.186	0.238
5	0	1	0	0	9	8.677	–0.444	0.254
6	0	1	0	1	6	8.186	–0.186	0.238
7	0	1	1	0	7	10.429	–0.186	0.238
8	0	1	1	1	44	41.569	0.096	0.274
9	1	0	0	0	1	1.458	–0.444	0.275
10	1	0	1	1	6	6.984	0.096	0.274
11	1	1	0	0	3	1.019	–0.186	0.238
12	1	1	0	1	3	4.063	0.096	0.274
13	1	1	1	0	12	5.174	0.096	0.274
14	1	1	1	1	141	144.017	0.651	0.521

Source: own calculations in R.

By default `factor.scores` function produces ability estimates for the observed response patterns. It also can be modified for non observed or specific response values.

4. Conclusions

In this paper we presented item response theory models. IRT present the relationship between the ability measures by the instrument and an item response. The IRF gives the probability that a person with a given ability level will answer correctly. Persons with lower ability have less of a chance, while persons with high ability are very likely to answer correctly; for example, students with higher math ability are more likely to get a math item correct. The exact value of the probability depends, in addition to ability, on a set of item parameters for the IRF.

We introduced models for dichotomous parameters such as the Rasch (1 PL) and the Birnbaum (2 and 3 PL) IRT models, as well as the theoretical four-parameters model. The basic key in IRT models is the model for the probability of a correct response in each item given the ability level. We presented estimation methods for

IRT models for item and individual latent score, we also described goodness of fit statistics and we applied IRT in R software.

In this paper we presented the application of item response theory models for the analysis of survey and social data based on data-set on abortion. We presented the R package for item response analysis and evaluating the Rasch model. The R package `ltm` provides a flexible framework for basic IRT analyses that covers some of the most common models for dichotomous and polytomous data. The main functions of the package have already been presented for different types of Rasch models with item characteristic curve and item information curve providing the relationship between a latent ability and the performance on a test item, and information about the ability of the examinee respectively.

Bibliography

Andrich D., 1978, *Application of a psychometric rating model to ordered categories which are scored with successive integers*, Applied Psychological Measurement, 2, pp. 581-594.

Bates D.M., 2008, *Fitting mixed-effects models using the lme4 package in R*, http://www.stat.wisc.edu/~bates/PotsdamGLMM/LMMD.pdf.

Barton M.A., Lord F.M., 1981, *An upper asymptote for the three-parameter logistic item-response model*, Princeton, NJ: Educational Testing Service.

Birnbaum A., 1957, *Efficient design and use of tests of a mental ability for various decision making problems*, Series Rep., no. 58-16, project no. 7755-23, Randolph Air Force Base, Tx: USAF School of Aviation Medicine.

Birnbaum A., 1958, *On the estimation of mental ability*, Series Rep., no. 15, project no. 7755-23, Randolph Air Force Base, Tx: USAF School of Aviation Medicine.

Birnbaum A., 1968, *Some Latent Trait Models*, [in:] Lord F.M., Novick M.R. (eds.), *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

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

Chalmers R.P., 2012, *mirt: A Multidimensional Item Response theory package for the R environment*, Journal of Statistical Software, 48 (6), pp. 1-29.

Chen W.-H., Thissen D., 1997, *Local independence indexes for item pairs using item response theory*, Journal of Education and Behavioral Statistics, 22, pp. 114-142.

DeBoeck P., Wilson M., 2004, *Explanatory Item Response Models*, Springer-Verlag, New York.

DeMars C., 2010, *Item Response Theory. Understanding Statistics Measurement*, Oxford University Press.

Doran H., 2010, *MiscPsycho: Miscellaneous Psychometric Analyses*, http://CRAN.R-project.org/package-MiscPsycho.

Doran H., Bates D., Bliese P., 2007, *Estimating the Multilevel Rasch model: With the lme4 Package*, Journal of Statistical Software, 20 (2), pp. 1-18.

Gulliksen H., 1950, *Theory of Mental Tests*, Harold Gulliksen, Wiley, London, Chapman & Hall, New York.

Hambleton R.K., Swaminathan H., 1985, *Item Response Theory: Principles and Applications*, Kluwer, Boston.

Hattie J., 1984, *An empirical study of various indices for detecting unidimensionality*, Multivariate Behavioral Research, 19, pp. 49-78.

Lord F.M., 1974, *Estimation of latent ability and item parameters when there are omitted responses*, Psychometrika, 39, pp. 247-264.
Lord F.M., 1974, *The relative efficiency of two tests as a function of ability level*, Psychometrika, 39, 351-358.
Lord F.M., Novick M.R., 1968, *Statistical Theories of Mental Test Scores*, Reading, Addison-Wesley.
Mair P., Hatzinger R., 2007, *Extended Rasch modeling: The eRm package for the application of IRT models in R*, Journal of Statistical Software, 20(9), pp. 1-20.
McDonald R.P., 1967, *Nonlinear factor analysis*, Psychometric Monographs 15, Chicago, University of Chicago Press.
McGrath K., Waterton J., 1986, *British Social Attitudes, 1983-86 Panel Survey*, SCPR, London.
Rasch G., 1960, *Probabilistic Models for Some Intelligence and Attainment Tests*, Danish Institute for Educational Research, Copenhagen, The University of Chicago Press, Chicago.
Reckase M.D., 2009, *Multidimensional Item Response Theory*, Springer-Verlag, New York.
Rizopoulos D., 2006, *ltm: An R package for latent variable modelling and item response theory*, Journal of Statistical Software, 17 (5), pp. 1-25.
Robitzsch A., Kiefer T., Wu M., 2017, *TAM: Test analysis modules. R package version 2*, https://CRAN.R-project.org/package=TAM, pp. 7-56, access: 3.11.2017.
Samejima F., 1969, *Estimation of latent ability using a response pattern of graded scores*, Psychometrika. Monograph Supplement, 34, 4.
Tate R., 2003, *A comparison of selected empirical methods for assessing the structure of responses to test items*, Applied Psychological Measurement, 27, pp. 159-203.
Van der Ark L.A., 2010, *Getting started with Mokken scale analysis in R,* unpublished manuscript, https://sites.google.com/a/tilburguniversity.edu/avdrark/mokken.
Weeks J., 2010, *plink: An R Package for Linking Mixed-Format Tests Using IRT-Based Methods*, Journal of Statistical Software, 35 (12), pp. 1-33.
Yen W.M., 1984, *Effects of local item dependence on the fit and equating performance of the three-parameter logistic model*, Applied Psychological Measurement, 8, pp. 125-145.

ANALIZA TEORII ODPOWIEDZI NA POZYCJE TESTOWE W TEORII POMIARU Z ZASTOSOWANIEM PAKIETU LTM PROGRAMU R

Streszczenie: Modele teorii odpowiedzi na pozycje testowe (modele IRT) są szczególnym rozszerzeniem klasycznej teorii testu (CCT). Modele te z powodzeniem wykorzystywane są w psychologii, medycynie, marketingu, a także w naukach społecznych. Modele teorii odpowiedzi na pozycje testowe opisują relację między obserwowalnymi cechami respondenta a nieobserwowalnymi zmiennymi lub zdolnościami poszczególnych osób odpowiadających na pytania. W niniejszym artykule zaprezentowano podstawowe modele IRT dla zmiennych niemetrycznych, m.in. model Rascha oraz Birnbauma. Przedstawiono również założenia, metodę estymacji, własności oraz procedury selekcji modeli. W niniejszym artykule wykorzystano program R oraz pakiet `ltm`, pozwalający na przeprowadzenie pełnej analizy opartej na modelach teorii odpowiedzi na pozycje testowe.

Słowa kluczowe: modele teorii odpowiedzi na pozycje, teoria pomiaru, analiza klas ukrytych, program R.