Social science measurement by means of item response models

Sijtsma, K.

Published in:
Proceedings in computational statistics

Publication date:
2000

Social science measurement by means of item response models

Klaas Sijtsma

Department of Methodology, FSW, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Abstract. The basic ideas of measurement in the social and behavioral sciences is explained, followed by a discussion of item response theory, which supplies the family of modern statistical measurement methods. Four specialized topics in item response modeling are discussed, that are at the core of present-day research in item response theory.

Keywords. goodness-of-fit methods, item response theory, item selection, misfitting respondents, person ordering

1 The basic ideas of social science measurement

The measurement of hypothetical constructs, such as intelligence, introversion, and achievement in psychology, and attitudes and opinions in sociology and political science, has a long and impressive tradition. Well-known names in psychological measurement are Spearman (classical test theory, factor analysis), Thurstone (paired comparison, factor analysis), and Lord (classical test theory, item response theory). In sociology, Guttman (scalogram analysis, covariance structures) and Lazarsfeld (latent class analysis, latent trait models) immediately come to mind. Except for Spearman, each of these milestone researchers has contributed to what is nowadays known as item response theory (Van der Linden & Hambleton, 1997).

Item response models are used for the analysis of data from psychological and educational tests and from questionnaires which are used throughout the behavioral and social sciences. The results of such data analyses can be used for assessing the composition of the final test or questionnaire, estimating the reliability and the validity of the measurement values collected by a test or a questionnaire, and assigning measurement values to individuals who took a test or filled out a questionnaire. Once in their final form, tests are used for diagnosis aimed at the assignment of individuals to treatment or therapy (counseling, clinical setting) and remedial teaching (education) or at giving job advice. Also, the decision to admit pupils to certain schools and select people in jobs is often supported by test performance results. Questionnaires are often used in such diverse areas as opinion polls, demographic research, and marketing research where the measurement of preferences may be of primary interest.

Tests and questionnaires almost always consist of a number (J) of problems (e.g., mazes, building blocks, geometric figures which have to be rotated mentally), questions (essay questions or open-ended questions, multiple-choice questions), or statements (e.g., about ethical or political issues or about a respondent's behavior). In general, these building blocks are called "items". The items reflect the final operationalization of the hypothetical construct

and are crucial to the success of measurement. The responses to items are assumed to contain information about a respondent's standing on the unobservable or latent trait measured by the complete set of items in the test or questionnaire.

Responses to items may be solutions to problems, choices, markings or written or oral reports. These responses are qualitative and as such they are not suited for statistical analysis. The necessary quantification is done by categorization of the responses, ordering of categories on the latent trait, and assigning ordered numbers to the categories which reflect the ordering on the latent trait. Thus, the quantification rests upon the idea: A higher item score reflects a higher standing on the latent trait. The statistical analysis by means of item response models of J item scores collected in a sample of N respondents has to point out whether the scoring rule made sense.

2 Item response models

Let X_j denote the random variable for the ordered score on item j ($j = 1, \ldots, J$), with realizations x_j, with $x_j = 0, 1, \ldots, m$. For $m \geq 2$, items are called polytomous items. Ordered polytomous scores may indicate the degree to which a respondent endorses an attitude statement such as "The present Dutch law on euthanasia is too liberal", but they may also indicate the rating of an essay on the exploitation of the tropical rain forests, written by high school students. Very common are dichotomous scores, $x_j = 0, 1$, where 0 indicates an incorrect answer and 1 a correct answer. Dichotomous scores are typical for multiple-choice items, but may also indicate agreement or disagreement with an attitude statement.

Furthermore, let $\mathbf{X}$ denote a vector with J item score variables and let $\mathbf{X} = \mathbf{x}$ denote the J realizations. Also, let θ denote the latent trait. We define $P(X_j = x|\theta)$ to be the conditional probability of obtaining a score of x on item j. This conditional probability describes the relationship of an item score to the latent trait, and is called the category response curve. For dichotomous items, this probability is defined as $P_j(\theta) \equiv P(X_j = 1|\theta)$, and is called the item response function. In general, we call category response curves and item response functions simply response functions.

In general, item response models rest on a number of assumptions. First, item response models can be multidimensional, which means that the items from a test simultaneously measure several traits, in which case θ is a vector, or unidimensional, in which case θ is a single latent variable. Most item response models are unidimensional, which reflects the common practice that tests and questionnaires are required to measure one construct at a time. Second, it is mostly assumed that respondents do not improve or otherwise change their ability, attitude, an so forth, while taking the test or filling out the questionnaire. That is, the measurement procedure does not affect the traits it attempts to measure (cf. a cold thermometer that, when placed in warmer water, affects the water temperature read from its scale; the idea is that this should be avoided). Sometimes, item response models drop the assumption of local independence, but this is rather uncommon.

For item scores $x_j = 0, \ldots, m$ and $m \geq 1$, the assumptions of unidimensionality and local independence can be written as

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{j=1}^{J} P(X_j = x_j|\theta).$$

By integrating θ out, we obtain the J-variate distribution of the item scores,

$$P(\mathbf{X} = \mathbf{x}) = \int_{\theta} \prod_{j=1}^{J} P(X_j = x_j|\theta) dG(\theta), \tag{1}$$

where $G(\theta)$ is the cumulative distribution function of θ. Without further restrictions on the response functions, the cumulative distribution of θ, or both, the multivariate distribution $P(\mathbf{X} = \mathbf{x})$ is not restricted in any way. This is undesirable because in order to be able to test whether an item response model fits empirical data, observable consequences need to be derived from the model. In item response modeling, this is accomplished by placing restrictions on the response functions.

Two kinds of restrictions are possible. *Parametric* item response models assume a particular parametric function. For example, for dichotomous items the much used 3-parameter logistic model (Lord & Novick, 1968; part IV by A.L. Birnbaum; Lord, 1980) assumes that

$$P(X_j = 1|\theta) = \gamma_j + \frac{(1 - \gamma_j) \exp[\alpha_j(\theta - \delta_j)]}{1 + \exp[\alpha_j(\theta - \delta_j)]}, \tag{2}$$

where δ_j is a parameter which locates the item on the θ scale, α_j is a parameter which is monotonically related to the slope of the function at δ_j, and γ_j is the lower asymptote for $\theta \to -\infty$. Parameter δ_j gives us an impression about the difficulty of the item, α_j shows how well the item separates low θs from high θs, and γ_j gives the probability of a correct answer for low θ (e.g., when low-ability respondents guess for the correct answer on a multiple-choice item). *Nonparametric* models only place order restrictions on response functions. For example, for dichotomous items the model of monotone homogeneity (Mokken, 1971) assumes that $P_j(\theta_a) \leq P_j(\theta_b)$ whenever $\theta_a < \theta_b$ (for all pairs of unequal θs), and any item response function that has this property is accepted. Both the 3-parameter logistic model and the model of monotone homogeneity express the idea that a higher θ should lead to a higher probability of answering an item correctly. For item scores expressing preferences (score 1 means stimulus was chosen, and score 0 otherwise), item response models assume that response functions are bell-shaped (Post, 1992).

The estimation of parametric item response models mostly is done with maximum likelihood methods, but Bayesian methods are also used (Baker, 1992; Patz & Junker, 1999). Because of the variety of parametric item response models and estimation methods, we only mention conditional maximum likelihood estimation that is used, in particular, for estimating the 1-parameter logistic model (obtained from Equation 2 by setting γ_j to 0 and α_j to 1; see Fischer & Molenaar, 1995). Here, the likelihood of the data $\mathbf{X}_{N \times J}$ given the marginal distribution for the items (columns) is maximized using iterative algorithms, such as the Newton-Raphson. This yields estimates of the δs, which are assumed fixed when in the next step the θs are estimated. For estimating the 2- and 3-parameter logistic models (the 2-parameter logistic model is obtained from Equation 2 by setting γ_j to 0; see Lord, 1980), marginal maximum likelihood estimation is used. Here, a distribution for θ is assumed and the likelihood is integrated across this distribution. Next, the likelihood is solved for the item parameters, which in the final step are assumed fixed when the θs are estimated. For nonparametric models, that do not define parametric response functions with latent variables, estimation is

relatively simple and can be done analytically, which usually does not pose any technical problems.

3 Some special topics

Due to space limitations, we will restrict ourselves to a brief discussion of four recent developments in item response theory, which together constitute the COMPSTAT 2000 symposium "Computer-intensive statistics in modern item response theory".

3.1 Goodness-of-fit methods

Usually, the univariate marginal distributions of the multivariate distribution of $\mathbf{X}$ (Equation 1) are used for investigating properties of response curves, for example, whether they comply with the shape described by Equation 2 (parametric) or whether they are nondecreasing functions of θ (nonparametric). The bivariate marginal distributions contain information about relationships between items and are used for evaluating the assumptions of unidimensionality and local independence. Other K-variate marginal distributions $(3 \leq K \leq J-1)$ contain information about simultaneous relationships between K items, but are almost never used for testing models due to the enormous complexity of the methods implied (Sijtsma & Junker, 1996).

There have been some attempts to reformulate item response models with continuous latent traits as latent class models with ordered latent classes; see, for example, Croon (1991). Here, the idea is that respondents do not have a score on a continuous θ, which indicates their level on that latent trait, but that someone with item score pattern $\mathbf{x}_i$ belongs to a discrete latent class q $(q = 1, \ldots, Q)$ with probability $P(q|\mathbf{x}_i)$. The number of ordered latent classes is limited and in this sense restricts the data structure. Van Onna (this symposium) formulated nonparametric item response models as ordered latent class models and constructed a Bayesian estimation algorithm for the model parameters, and also proposed goodness-of-fit methods.

3.2 Automated item selection from an item pool

In the context of nonparametric item response theory, three automated item selection procedures have been developed (similar procedures do not yet exist for parametric models). First, in the context of the model of monotone homogeneity a sequential clustering algorithm has been designed that starts with the subset of items (mostly, two items) from a pool of I items that best complies with the selection criterion. In each next step, an item is selected that together with the already selected items maximizes the selection criterion. This continues until all items are selected or no more items can be added that satisfy auxiliary selection criteria. The second procedure selects items stepwise according to a hierarchical clustering technique. In the first step, the best pair according to a particular formal criterion is selected; in the second step, either a third item is added to this pair or a second pair is selected; in the third step, either an item is added to the triplet already selected, to one of the two pairs already selected, or a third pair is selected; and this continues until finally all items are selected together in one set of J items. The researcher has to decide which intermediate clustering solution is his/her final solution. The third procedure searches for the subset of J items $(J = 2, 3, 4, \ldots, I)$ from the pool of I items that maximizes a particular formal criterion. Theoretically, this procedure tries all possibilities and may form several non-overlapping scales. In practice, algorithms have been designed to

speed up the search from the enormous amount of possible scales. Van Abswoude (this symposium) recently has started a systematic comparison of the three item selection procedures. This comparison entails the effectiveness of each procedure for finding from simulated data with a known dimensionality structure the correct dimensionality and the correct assignment of items to subscales.

3.3 Using X_+ to order persons on θ

Psychologists and measurement practitioners may test children and adults, report the test results to these people, but perhaps also to school, companies and other clients. In all cases, professionals using tests and questionnaires often like to use and report simple scores, such as the number-correct or the number of points earned on a questionnaire, to be denoted $X_+ = \sum X_j$. When measurement instruments are used for ordering respondents, the ordering on X_+ should in a stochastic way reflect the ordering on θ. That is, the conditional cumulative distributions of θ should be higher for higher values of X_+: For integers $s_1 < s_2$, and for arbitrary real values t, we have that

$$P(\theta \leq t | X_+ = s_1) \geq P(\theta \leq t | X_+ = s_2). \tag{3}$$

For dichotomous items, Equation 3 is implied by unidimensional, locally independent item response models with monotonely nondecreasing item response functions. Most of the well known polytomous item response models, *theoretically* do not imply Equation 3 (Hemker, Sijtsma, Molenaar, & Junker, 1997). For three known classes of polytomous item response models, however, Van der Ark (this symposium) found in simulated *practically* relevant situations that for partial credit models and sequential ratio models Equation 3 was almost never violated, but that for graded response models a nonignorable number of violations occurred.

3.4 Adaptive test procedures and misfitting respondents

Item response theory has facilitated adaptive testing, a computerized testing procedure that adapts the difficulty level of the items stepwise to the estimated ability level of the examinee (Van der Linden & Glas, 1999). For example, after an examinee has received the first five items (presented on a computer screen; the examinee responds by pushing a button) the adaptive testing software selects the sixth item on the basis of the intermediate estimate of θ. The difficulty level of the sixth item corresponds to the estimated θ value. The result of this adaptive testing procedure is a highly reliable estimated θ, measured with relatively few items. This testing procedure is important, in particular, in educational measurement, where huge numbers of items are needed for testing large numbers of pupils. The items are stored in a so-called item bank, from which the adaptive testing software can select large numbers of different or partly overlapping tests.

Possible problems during adaptive testing are, for example, that an individual examinee may have pre-knowledge of some of the more difficult items from the item bank and when presented with such items, obtains better results than when presented with other items of equal difficulty, but previously unknown to him/her. Also, due to extreme anxiety the examinee may fail on the first series of moderate to easy items, and only catch up on the real θ level later on. Meijer (this symposium) devised a method for detecting unusual item score patterns at the individual level, that accumulates consecutive deviation scores that are larger than a preset level, and that identifies an item score pattern as aberrant when the index is higher than a particular

critical value. The idea comes from industrial quality control and has proved to be useful in computerized adaptive testing.

References

Baker, F.B. (1992). *Item Response Theory. Parameter Estimation Techniques.* New York: Marcel Dekker.

Croon, M.A. (1991). Investigating Mokken scalability of dichotomous items by means of ordinal latent class analysis. *British Journal of Mathematical and Statistical Psychology*, **44**, 315-331.

Fischer, G.H., and Molenaar, I.W. (1995). *Rasch Models. Foundations, Recent Developments, and Applications.* New York: Springer-Verlag.

Hemker, B.T., Sijtsma, K., Molenaar, I.W., and Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, **62**, 331-347.

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

Lord, F.M., and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores.* Reading, MA: Addison-Wesley.

Mokken., R.J. (1971). *A Theory and Procedure of Scale Analysis.* Berlin: De Gruyter.

Patz, R.J., and Junker, B.W. (1999). A straightforward approach to Markov Chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, **24**, 146-178.

Post, W.J. (1992). *Nonparametric Unfolding Models. A Latent Structure Approach.* Leiden: DSWO Press.

Sijtsma, K., and Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology*, **49**, 79-105.

Van der Linden, W.L., and Glas, C.A.W. (1999). *Computer Adaptive Testing: Theory and Practice.* Dordrecht, the Netherlands: Kluwer.

Van der Linden, W.J., and Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory.* New York: Springer-Verlag.