September, 2014

Evaluation of item parameter recovery estimation by ACER ConQuest software

Luc T Le, *ACER*
Ray Adams, *ACER*

Evaluation of item parameter recovery estimation by ACER ConQuest software

Luc T. Le & Raymond Adams
Australian Council for Educational Research
luc.le@acer.edu.au

Proposed networks: (9) Assessment, Evaluation, Testing and Measurement

Intent of publication: European Educational Research Journal, Journal of Educational Measurement, or International Journal of Testing.

Key words: testing, assessment, Rasch model, estimation method, ConQuest software

RATIONAL

ACER ConQuest (Adams, Wu, and Wilson, 2012) has been popularly used for analysing testing and assessment data. Two of the most common estimation methods for Rasch measurement models (Rasch, 1960/1980) are available in this software, marginal maximum likelihood estimation (MML) and joint maximum likelihood estimation (JML).

Under the JML method, as developed by Birnbaum (1968), and Wright and Pachapakesan (1969), all item parameters and all person parameters are regarded as fixed unknowns to be estimated. Therefore, the parameters involved in the estimation procedure of this method are all of the case parameters and all of the item parameters.

The MML method was developed by Bock and Lieberman (1970), and Bock and Aitken (1981). Under this method, item parameters are considered as “structural”, while ability parameters are “incidental”. It is assumed that individual’s positions on the latent variable are sampled from distributions of possible values. As a result, in its estimation procedure, the MML includes the item parameters and the population parameters but not case parameters. Although the distribution can be of any type, with a limit on the number of parameters, normal densities are most frequently used (see Mislevy, 1984). Additionally, it can be applied with a discrete distribution where a fixed set of grid points is assumed and a weight is estimated at each grid point (Adams and Wilson, 1996).

From a theoretical perspective however JML has some shortcomings. Andersen (1973) showed that JML estimates of the item parameters for Rasch models are not consistent if the number of items is fixed and the size $N \rightarrow \infty$. To deal with the bias in

JML, Wright and Douglas (1978) proposed a correction of, (K-1)/K, where K is the number of items. They argued that this correction removed most of the bias for K>20. For tests of fewer than 15 items, van den Wollenberg, Wierda, and Jansen (1998) suggested that this bias correction is inappropriate since the bias was dependent not only on the number of items, but also on the skewness of the item difficulty distribution.

A second potential shortcoming of JML is that in many of its potential applications the goal is to make inferences concerning populations. If JML is used for estimating the measurement model then a two-step analysis is required. First the case parameters are all estimated with JML and then the population parameters are estimated from individual case estimates. A number of researchers have illustrated that the use of case parameter estimates as though they were true values in a two-step analysis can lead to quite misleading outcomes.

The MML method can overcome these disadvantages of the JML method. Particularly, if both the item response models and the assumed population distributions are correct the MML item parameter estimates are consistent (Bock and Aitkin, 1981). Additionally, population parameters are estimated directly from the observed responses to avoid the problems associated with estimating population characteristics using fallible case parameter estimates in a two-step process.

However, the application of MML approach is often restricted to the assumption of a distribution for the population when this may not be a desirable assumption. Some empirical studies demonstrate that MML estimators loose accuracy and efficiency when the prior assumption of the latent distribution is violated (Yen, 1987; Drasgow, 1989; Seong, 1990; Harwell and Janosky, 1991; Stone, 1992).

This study is concerned with item parameter recovery for the dichotomous Rasch model. Our primary focus is on comparing JML and MML when the assumptions of MML are violated, that is the abilities are not sampled from the distribution that is assumed in the estimation.

STUDY METHODOLOGY

We generate data that conforms to the dichotomous Rasch model using four alternative *true* population distributions for the abilities. We then use the ACER

ConQuest software to recover Rasch model parameter estimates using JML and MML. For the MML estimation we consider two alternative distribution assumptions. First, we assume a normal population distribution, the variance of which is estimated, this will be referred to as MML-Normal. Second, we assume a discrete population distribution, under which a set of nodes uniformly spaced in a certain interval is assumed and densities at each node are estimated, this will be referred to as MML-Discrete.

The accuracy of parameter recovery is shown by computing bias and root mean square error (RMSE) statistics for each of the estimated parameters. Bias for an item difficulty parameter or the variance parameter was computed as the mean difference, across the replications, between the estimated values and the true values.

DATA

For the simulation study a number of factors that can be varied need to be considered. The characteristics of the population distribution, the size of the ability sample, the characteristics of the item distribution and the length of the tests. For the sake of simplicity and to ensure focus on the shape of the population distribution, eight distinct combinations of the above listed factors were considered – four population distributions (to be described below), a single sample size of 2000 examinees, a single uniform U[–3,3] item distribution and two test lengths (10 and 50 items). The item difficulties of 10 and 50 items were randomly generated from a uniform distribution U[–3,3] and then transformed to ensure constrained as a mean of zero. These values were fixed and considered as the generated values for all replications. For each of the eight combinations of factors 1000 replications was undertaken.

The central variable in this investigation was the shape of the population distribution. The four distributions used in this study are normal, bimodal, uniform and chi-square. For comparison purposes all four distributions had a mean of zero and standard deviation of one. The normal distribution was $N(0,1)$. The uniform distribution was $U[-\sqrt{3},+\sqrt{3}]$. The bimodal distribution was a combination of two normal distributions with means of –0.8 and 0.8 respectively, and standard deviation of $\sqrt{0.6}$, $N\left(-0.8,\sqrt{0.6}\right)$and $N\left(+0.8,\sqrt{0.6}\right)$. The chi-square distribution was a standardisation of a chi-square distribution with five degree of freedom.

RESULTS

In general, results in this study are consistent with the findings from a number of previous studies (e.g., Yen, 1987; Drasgow, 1989; Harwell and Janosky, 1991; Stone, 1992). As expected, results showed that the accuracy of JML was dependent on test length. MML-Normal was the best method when the assumption of ability distribution was matched. However, the accuracy or MML-Normal decreased with the violation level of the assumption of normal distribution of the latent ability. The MML-Discrete estimation could overcome well the weakness of the MML-Normal when the normality of the distribution was violated.

Particularly, with a 50-item test, the three methods tended to produce similar results with small or negligible bias in item parameter estimates, although MML-Normal provided more accurate estimates than JML and MML-Discrete when the assumption of ability distribution was matched. With 10-item test, while JML estimates showed large bias, MML-Normal still provided very reliable estimates in a 10-item test when the assumption of ability distribution was matched. However, this method appeared to produce the large bias when the ability distribution was skewed.

REFERENCE

Adams, R. J., and Wilson, M. R. (1996). A random coefficients multinomial logit: A generalized approach to fitting Rasch models. In *Objective Measurement III: Theory into Practice*, G. Engelhard and M. Wilson (eds.), pp 143–166. Norwood, New Jersey: Ablex.

Adams, R.J., Wu, M.L., & Wilson, M.R. (2012) ACER ConQuest 3.0. [computer program]. Melbourne: ACER.

Andersen, E. B. (1973). Conditional inference in multiple choice questionnaires. *British Journal of Mathematical and Statistical Psychology*, *26*, 31–44.

Birnbaum, A. (1968). Some Latent trait models and their use in inferring an examinee’s ability. In F. M. Lord and M. R. Novick (Eds*.) Statistical Theories of Mental Scores* (pp. 397–472). Reading, MA: Addition-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, *46*, 443–459.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for dichotomously scored items. *Psychometrika*, *35*, 179–187.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement*, *13*, 77–90.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement*, *15*, 279–291.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*, 359–381.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* (expanded ed). Chicago. The University of Chicago Press. (Original work published 1960).

Seong, T. (1990). Sensitivity of Marginal Maximum Likelihood Estimation of Item and Ability Parameters to the Characteristics of the Prior Ability Distributions. *Applied Psychological Measurement* , *14*, 299–311.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1–16.

van den Wollenberg, A. L., Wierda, F. W., & Jansen, P. G. W. (1998). Consistency of Rasch Model Parameter Estimation: A Simulation Study. *Applied Psychological Measurement*, *12,* 307–313.

Wright, B. D., & Panchapakesan, N. A. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, *29*, 23–48.

Wright, B. D., & Douglas, G. A. (1978). Better procedures for sample-free item analysis. *MESA Memorandum No. 20.* Chicago, IL: University of Chicago, Department of Education.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika*, *52*, 275–291.