***irtplay*: An R Package for Online Item Calibration, Scoring, Evaluation of Model Fit, and Useful Functions for Unidimensional IRT**

Hwanggyu Lim¹ and Craig S. Wells²

Abstract

The R package *irtplay* provides practical tools for unidimensional item response theory (IRT) models that conveniently enable users to conduct many analyses related to IRT. For example, the *irtplay* includes functions for calibrating online items, scoring test-takers' proficiencies, evaluating IRT model-data fit, and importing item and/or proficiency parameter estimates from the output of popular IRT software. In addition, the *irtplay* package supports mixed-item formats consisting of dichotomous and polytomous items.

Keywords

computerized adaptive testing, IRT model fit, item response theory, scoring

Description

The R package *irtplay* provides practical tools for unidimensional item response theory (IRT) models that conveniently enable users to conduct many analyses related to IRT such as calibrating online items, scoring test-takers' proficiencies, and evaluating IRT model-data fit. The *irtplay* software package is quite user-friendly and its functions are easy to implement. In addition, the *irtplay* package supports mixed-item formats consisting of dichotomous and polytomous items. For the polytomous items, the graded response model and the (generalized) partial credit models are available.

Online Item Calibration Using Stocking's Method A

In computerized adaptive testing (CAT), online items, which include pretest items and operational items, should be calibrated precisely to ensure a healthy item bank (Stocking, 1988). The *irtplay* package calibrates the online items using Stocking's (1988) Method A, which is the maximum likelihood estimation (MLE) of the item parameters given the proficiency estimates. Method A is known to result in accurate, unbiased item parameter calibration when items are randomly administered to examinees (Chen & Wang, 2016; He et al., 2017). Although Method

¹Graduate Management Admission Council, Reston, VA, USA

²University of Massachusetts Amherst, USA

Corresponding Author:

Hwanggyu Lim, Graduate Management Admission Council, 11921 Freedom Drive, Suite 300, Reston, VA 20190, USA.
Email: hlim@gmac.com

A is the simplest and most straightforward online calibration method in current use, there is little software available to support it. The updated version of *irtplay* will include fixed item parameter calibration (FIPC; Kim, 2006).

Estimation of Proficiency Parameters

For estimation of proficiency parameters, six popular scoring methods can be implemented using the *irtplay* package. Available methods include MLE, MLE with fences (MLEF; Han, 2016), maximum a posteriori (MAP) estimation (Hambleton et al., 1991), expected a posteriori (EAP) estimation (Bock & Mislevy, 1982), EAP summed scoring (Thissen & Orlando, 2001), and inverse test characteristic curve scoring (Kolen & Tong, 2010). Although much IRT software has been developed, there are few software packages available to estimate proficiency parameter with the mixed-item format.

Evaluation of IRT Model-Data Fit on Item Level

In terms of assessing the IRT model-data fit, the *irtplay* package supports computation of popular item-fit statistics. Available fit statistics include χ^2 fit statistic (Bock, 1960; Yen, 1981), likelihood ratio χ^2 fit statistic (G2; McKinley & Mills, 1985), infit and outfit statistics (Ames & Penfield, 2015), and $S\text{-}\chi^2$ (Orlando & Thissen, 2000). In addition, *irtplay* provides users with graphical displays to examine residuals between the observed data and model-based prediction for items (Hambleton et al., 1991).

Useful Functions Related to IRT

The *irtplay* package also includes many other convenient functions such as importing item and/or proficiency parameter estimates from the output of popular IRT software—Bilog-MG, PARSCALE, flexMIRT, and *mirt* (R package), for example. The software also enables users to implement flexMIRT software through R console, generate simulated item response data, compute asymptotic variance–covariance matrices of item parameter estimates (Li & Lissitz, 2004), compute conditional distribution of observed scores using the Lord and Wingersky (1984) recursion formula, compute item and test information functions, compute item and test characteristic curves, and plot item and test characteristic curves and item and test information functions.

Availability and Distribution

The R package *irtplay* and its reference manual are available for free from the website <http://www.r-project.org> and works on Windows, Linux, and MacOS platforms. Version 3.6.0 (or later) of the R software should be installed for optimal functioning of the *irtplay* package.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Hwanggyu Lim <https://orcid.org/0000-0002-5714-0122>

Supplemental Material

Supplementary material is available for this article online (<https://cran.r-project.org/web/packages/irtplay/irtplay.pdf>)

References

- Ames, A. J., & Penfield, R. D. (2015). An NCME instructional module on item-fit statistics for item response theory models. *Educational Measurement: Issues and Practice*, 34(3), 39–48.
- Bock, R. D. (1960). *Methods and applications of optimal scaling*. L. L. Thurstone Psychometric Laboratory.
- Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Psychometrika*, 35, 179–198.
- Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika*, 81(3), 674–701.
- Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE.
- Han, K. T. (2016). Maximum likelihood score estimation method with fences for short-length tests and computerized adaptive tests. *Applied Psychological Measurement*, 40(4), 289–301.
- He, Y., Chen, P., Li, Y., & Zhang, S. (2017). A new online calibration method based on Lord's bias-correction. *Applied Psychological Measurement*, 41(6), 456–471.
- Kim, S. (2006). A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43(4), 355–381.
- Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, 29(3), 8–14.
- Li, Y., & Lissitz, R. (2004). Applications of the analytically derived asymptotic standard errors of item response theory item parameter estimates. *Journal of Educational Measurement*, 41(2), 85–117.
- Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8(4), 453–461.
- McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, 9, 49–57.
- Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64.
- Stocking, M. L. (1988). *Scale drift in on-line calibration* (Research Report No. 88-28). Educational Testing Service.
- Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Lawrence Erlbaum.
- Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, 5, 245–262.