Measurement: Interdisciplinary Research and Perspectives

ISSN: 1536-6367 (Print) 1536-6359 (Online) Journal homepage: https://www.tandfonline.com/loi/hmes20

R Packages for Item Response Theory Analysis: Descriptions and Features

Youn-Jeng Choi & Abdullah Asilkalkan

MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
2019, VOL. 17, NO. 3, 168–175
https://doi.org/10.1080/15366367.2019.1586404

SOFTWARE REVIEW

R Packages for Item Response Theory Analysis: Descriptions and Features

Youn-Jeng Choi and Abdullah Asilkalkan

Department of Educational Studies in Psychology, Research Methodology & Counseling, University of Alabama

ABSTRACT
About 45 R packages to analyze data using item response theory (IRT) have been developed over the last decade. This article introduces these 45 R packages with their descriptions and features. It also describes possible advanced IRT models using R packages, as well as dichotomous and polytomous IRT models, and R packages that contain applications such as differential item functioning and equating are also introduced. Thus, this article helps researchers who plan to use IRT-based analysis to decide on the type of IRT analysis and choose the appropriate R packages.

KEYWORDS
R package; item response theory; differential item functioning; equating; simulation study

Item response theory (IRT) is the most widely used measurement theory. Unlike classical measurement models, IRT models include these desirable features: item and test characteristics are group independent, examinee scores are not test dependent, item analysis is performed at the item level rather than at the test level, and a measure of precision for each ability score is provided (Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991). To date, data analysis using IRT has been supported by commercial IRT computer software such as BILOG-MG, flexMIRT, IRTPRO, Mplus, and PARSCALE. However, this software has barriers to researchers, especially those who are not familiar with IRT, as the manuals are long and hard to understand, and the software is not free. The need for better IRT applications has led to the development of R packages using IRT. About 45 R packages for IRT-based analysis have been developed over the last decade.

The purpose of this paper is to introduce R packages using IRT and to explain what types of measurement analysis can be possible using these R packages. Thus, this paper helps researchers who plan to use IRT-based analysis to decide on the type of IRT analysis and choose the appropriate R packages.

Tables 1–3 contain summaries of IRT R packages including dichotomous IRT models, polytomous IRT models, parameter estimation methods, goodness of fit, plots, information function for both item and test, and simulation study. As well as the estimation methods introduced in Table 2, the plRASCH package provides pseudolikelihood estimations as well as MLE on Rasch family models for polytomous items and multiple latent variables. Two R packages draw diagrams for IRT analysis: the WrightMap package contains powerful graphics tools such as item maps or item-person maps, which is used to describe unidimensional and multidimensional responses (Li & Hong, 2014; Torres Irribarra & Freund, 2014), and the latdiag package provides graphs for checking for latent scales for Guttman scale, Rasch scale, and Mokken double monotone scale (Dewey, 2018).

Five R packages provide the diagnostic tests to check IRT assumptions: unidimensionality, local independence, and homogenous population. For unidimensionality, there is the ltm package using modified parallel analysis, the eRm package using decreased inter-item correlations from Ponocny (2001)'s *T* statistics, and the pcIRT package using a likelihood ratio test. For local independence,

CONTACT Youn-Jeng Choi ychoi26@ua.edu Department of Educational Studies in Psychology, Research Methodology & Counseling, The University of Alabama, 302 Carmichael Hall, Tuscaloosa, AL 35487.

Table 1. Dichotomous and polytomous IRT models in R packages.

R package	Reference	Rasch	2PL	3PL	4PL	PC	RS	GPC	GR	NR
cacIRT	Lathrop (2015)	✓	✓	✓						
catR	Magis and Raîche (2012)	✓	✓	✓	✓	✓	✓	✓	✓	✓
CDM	George, Robitzsch, Kiefer, Groß, and Ünlü (2016)	✓	✓			✓		✓		
classify	Wheadon (2014)	✓	✓	✓		✓		✓		
DFIT	Cervantes (2017)	✓	✓	✓		✓		✓	✓	
difNLR	Drabinova et al. (2018)	✓	✓	✓	✓					
difR	Magis et al. (2010)	✓	✓	✓						
eRm	Mair and Hatzinger (2007)	✓				✓				
EstCRM	Zopluoglu (2015)									
FLIRT	Jeon, Rijmen, and Rabe-Hesketh (2014)	✓	✓	✓		✓			✓	
immer	Robitzsch and Steinfeld (2018)	✓	✓			✓	✓			
IRTpp	SICS Research Group (2016)	✓	✓	✓						
irtProb	Raiche (2014)	✓	✓	✓	✓					
irtoys	Partchev et al. (2017)	✓	✓	✓						
kequate	Andersson et al. (2013)		✓	✓				✓	✓	
LNIRT	Fox et al. (2018)									
lordif	Choi et al. (2011)							✓	✓	
ltm	Rizopoulos (2006)	✓	✓					✓	✓	
mirt	Chalmers (2012)	✓	✓	✓	✓	✓	✓	✓	✓	✓
mirtCAT	Chalmers (2016)	✓	✓	✓	✓	✓	✓	✓	✓	✓
mixRasch	Willse (2014)	✓				✓	✓			
MLCIRTwithin	Bartolucci and Bacci (2016)	✓	✓			✓			✓	
mRm	Preinerstorfer (2016)	✓								
MultiLCIRT	Bartolucci et al. (2014)	✓	✓			✓	✓	✓	✓	✓
pairwise	Heine (2019)	✓				✓				
pcIRT	Hohensinn (2018)	✓					✓			
PP	Reif (2017)	✓	✓	✓	✓	✓		✓		
psychomix	Frick et al. (2012)	✓								
SNSequate	González (2014)	✓		✓						
TAM	Robitzsch et al. (2018)	✓	✓	✓		✓	✓	✓		✓
xxIRT	Luo (2018)			✓				✓		

Note. PC = partial credit, RS = rating scale, GPC = generalized partial credit, GR = graded response, NR = nominal response.

there is the eRm package using increased inter-item correlations from Ponocny's *T* statistics, the Mokken package using conditional association, and the TAM package using residuals. Finally, for homogeneous population, there is the pcIRT package using Andersen's likelihood ratio test, which splits the data set into subsamples to test people's homogeneity (Hohensinn, 2018; Mair & Hatzinger, 2007; Rizopoulos, 2006; Robitzsch, Kiefer, & Wu, 2018; van der Ark, 2007).

Advanced IRT models

The violation of IRT assumptions (i.e., unidimensionality, local independence, homogeneous population, and non-speededness) can be solved by applying advanced IRT models. When the test and/or latent ability is not unidimensional, multidimensional IRT models can be applied. Ten R packages provide estimation using multidimensional IRT models: the FLIRT package for uni- and multidimensional explanatory IRT models, the irtProb package for multidimensional person IRT models, the mirtCAT package for multidimensional computerized adaptive tests (CAT) using IRT methodology, the MLCIRTwithin package for latent class IRT models under within-item multidimensionality, the MultiLCIRT package for multidimensional latent class IRT models, the pcIRT package for multidimensional polytomous Rasch model (Rasch, 1961) and continuous rating scale model by Mueller (1987), and the CDM, IRTpp, mirt, and TAM packages for uni- and multidimensional IRT models (Bartolucci & Bacci, 2016; Bartolucci, Bacci, & Gnaldi, 2014; Chalmers, 2016, 2012; George, Robitzsch, Kiefer, Groß, & Ünlü, 2016; Hohensinn, 2018; Jeon, Rijmen, & Rabe-Hesketh, 2014; Raiche, 2014; Robitzsch et al., 2018; SICS Research Group, 2016).

Table 2. Estimation methods in R packages.

R packages	MLE	CMLE	JMLE	MMLE	WLE	EAP	MAP	MCMC
cacIRT	✓							
catR	✓				✓	✓		
CDM	✓		✓	✓	✓	✓	✓	
classify								
DFIT								
difNLR	✓							
difR				✓				
eRm		✓	✓					
EstCRM	✓			✓				
FLIRT	✓					✓		
immer	✓	✓	✓					✓
IRTpp						✓	✓	
irtProb	✓						✓	
irtoys	✓				✓	✓		
kequate								
LNIRT								✓
lordif						✓		
ltm	✓					✓		
mirt	✓				✓	✓	✓	
mirtCAT	✓				✓	✓	✓	
mixRasch			✓					
MLCIRTwithin	✓							
mRm		✓						
MultiLCIRT	✓							
pairwise					✓			
pcIRT	✓	✓						
PP	✓				✓	✓	✓	
psychomix	✓	✓						
SNSequate	✓					✓	✓	
TAM	✓		✓	✓	✓	✓		
xxIRT			✓	✓		✓		

MLE = maximum likelihood estimation, JMLE = joint MLE, MMLE = marginal MLE, WLE = weighted likelihood estimation, EAP = expected a posteriori estimation, MAP = maximum a posteriori estimation, MCMC = Markov chain Monte Carlo.

The content of an item must not provide clues to the answers of other test items under the assumption of local independence. When a test contains a testlet format, an examinee's responses within the testlet are related to each other and thus not independent. Two R packages (the TAM for the Rasch testlet model using a bi-factor model and the mirt package for dichotomous and polytomous responses using bi-factor and two-tier analyses) provide analysis using the testlet IRT models, which control the testlet effects (Chalmers, 2012; Robitzsch et al., 2018).

When a population is not homogeneous and contains discrete latent classes, mixture IRT models can be applied using five R packages: the Psychomix package for the mixture Rasch model using CMLE, the Bradley–Terry mixture models, and multinomial processing tree mixture models (Frick, Strobl, Leisch, & Zeileis, 2012); the mirt package for latent class models such as multidimensional latent class and mixture IRT models (Chalmers, 2012); the mixRasch package for the dichotomous Rasch model, the rating scale model, and the partial credit model using JMLE (Willse, 2014); the mRm package for the dichotomous Rasch model using CMLE (Preinerstorfer, 2016); and the TAM package for mixture IRT models using MMLE (Robitzsch et al., 2018). For a speededness test, one can try the LNIRT package using the log-normal response time IRT model, which allows the analysis of both responses and response times with a Markov chain Monte Carlo (MCMC) Bayesian algorithm (Fox, Klotzke, & Entink, 2018).

The Irtrees package provides tree-based IRT models of the generalized linear mixed models family: linear response tree models (e.g., missing response models), nested response tree models, linear latent-variable tree models (e.g., models for change processes), and nested latent-variable tree models (e.g., bi-factor models), while the FLIRT package allows the generalized item response tree model (De Boeck & Partchev, 2012; Jeon et al., 2014).

Table 3. R Packages Providing Graphics, Goodness of Fit, and Simulation Study

	Graphics		Goodness of fit		Simulation study		
R packages	Plot	Information function	X^2	AIC & BIC	Data generation	R code example	Recovery analysis
cacIRT		✓			✓		
catR	✓	✓			✓	✓	
CDM	✓		✓	✓	✓	✓	
classify	✓				✓	✓	
DIFlasso	✓			✓	✓		
difNLR	✓			✓	✓		
emIRT					✓	✓	
eRm	✓	✓	✓	✓	✓		
EstCRM	✓		✓		✓	✓	✓
FLIRT	✓	✓		✓			
immer	✓				✓	✓	
IRTpp	✓	✓		✓	✓	✓	
irtProb	✓			✓	✓	✓	
irtoys	✓	✓	✓		✓		
kcirt					✓		
kequate	✓				✓		
LNIRT	✓				✓		
lordif	✓		✓		✓		
ltm	✓	✓	✓	✓	✓		
mirt	✓	✓	✓	✓	✓	✓	
mirtCAT	✓				✓	✓	
mixRasch	✓			✓	✓		
mRm	✓			✓	✓		
mstR					✓		
mudfold	✓				✓	✓	
MultiLCIRT	✓			✓	✓		
pairwise	✓	✓		✓	✓		
pcIRT	✓		✓		✓		
plRASCH					✓		
PP			✓		✓	✓	
psychomix	✓			✓	✓		
SNSequate	✓		✓		✓		
TAM	✓	✓	✓	✓	✓	✓	
WrightMap					✓		
xxIRT	✓				✓	✓	

The hierarchical IRT models can be analyzed using the emIRT and immer packages. The emIRT package can be used for the binary responses and the immer package for multiple ratings using the hierarchical rater model (Patz, Junker, Johnson, & Mariano, 2002) with MCMC estimation using a Metropolis–Hastings algorithm (Imai, Lo, & Olmsted, 2017; Robitzsch & Steinfeld, 2018).

One can also select R packages based on the type of responses. The kcirt package allows users to analyze the ipsative responses in forced-choice questionnaires using k-cube Thurstonian IRT models (Brown & Maydeu-Olivares, 2013; Zes, Lewis, & Landis, 2014). Continuous responses can be analyzed with the EstCRM package using Samejimas's continuous response model via MMLE and EM algorithm, and the pcIRT package can be used for the rating scale model for continuous data by Mueller (1987) (Hohensinn, 2018; Mueller, 1987; Samejima, 1973; Zopluoglu, 2015).

The nonparametric IRT models can be analyzed using six R packages: eRm, fwdmsa, KernSmoothIRT, Mokken, mudfold, and RaschSampler. The eRm and RaschSampler packages provide nonparametric Rasch model tests, and the KernSmoothIRT package performs nonparametric IRT models using kernel smoothing techniques (Mair & Hatzinger, 2007; Mair, Hatzinger, & Verhelst, 2015; Mazza, Punzo, & McGuire, 2014). The Mokken scale analysis (MSA; Mokken, 1971; Sijtsma & Molenaar, 2002; Sijtsma & Van der Ark, 2017), one of the more popular nonparametric IRT models, is used by the fwdmsa and Mokken packages, and the Mokken package provides the forward search for outlier detection in MSA (van der Ark, 2007;

Zijlstra, 2011). The mudfold package is for a nonparametric probabilistic model for unidimensional unfolding to analyze proximity items presumably generated from a nonmonotonic (unimodal) item response function (Balafas, Krijnen, Post, & Wi, 2017; Post, 1992; Van Schuur, 1984).

Differential item functioning (DIF)

There are seven packages to contain DIF using IRT or test scores from IRT estimation: CDM, DFIT, DIFlasso, difNLR, difR, lordif, and mirt. The CDM package detects both uniform and nonuniform DIF in the generalized DINA model by using the Wald test when a multiple group GDINA model is previously fitted (George et al., 2016). The DFIT package is useful for detecting DIF using Raju's area measures and ETS Delta measure for Mantel–Haneszel DIF statistics (Cervantes, 2017). It includes Monte Carlo item parameter replication method to obtain the associated statistical significance tests for cut-off points and power calculation (Oshima, Raju, & Nanda, 2006). The DIFlasso package contains DIF in Rasch models using a penalty approach (Schauberger, 2017).

The difNLR package contains methods for detection of DIF among dichotomous items and differential distractor functioning (DDF) among unscored items (Drabinova, Martinkova, & Zvara, 2018). Both uniform and non-uniform DIF effects can be detected using a likelihood-ratio test or *F*-test of the sub-model. The DDF is detected by using a multinomial log-linear regression model. The difR package contains several traditional methods for detecting DIF in dichotomous items (Magis, Beland, Tuerlinckx, & De Boeck, 2010). Both uniform and non-uniform DIF effects can be detected, with methods based on IRT methods or test score. Some methods deal with multiple group DIF. Methods available are (1) transformed item difficulties method, (2) Mantel–Haenszel, (3) standardization, (4) Breslow–Day, (5) logistic regression, (6) SIBTEST and Crossing-SIBTEST, (7) Lord's chi-square test, (8) Raju's area, (9) likelihood-ratio test, (10) generalized Mantel–Haenszel, (11) generalized logistic regression, and (12) generalized Lord's chi-square test.

The lordif package performs a logistic ordinal regression DIF using IRT for dichotomous and polytomous items (Choi, Gibbons, & Crane, 2011). The graded response model or the generalized partial credit model is used for IRT ability estimation. The mirt package for multidimensional IRT can also detect DIF using the Wald and likelihood-ratio approaches. It performs various omnibus differential item (DIF), bundle (DBF), and test (DTF) functioning procedures with computation of effect size measures of DIF (Chalmers, 2012).

Equating

Six R packages (ltm, equateIRT, kequate, SNSequate, irtoys, and TAM) provide different types of equating methods. The Ltm package provides the common item equating when common and unique items are analyzed simultaneously or when different sets of unique items are analyzed separately based on previously calibrated anchor items (Rizopoulos, 2006). The equateIRT package contains the computation of direct, chain, and average equating coefficients with standard errors using IRT methods for dichotomous items (Battauz, 2015). Test scoring can be equated by true score equating and observed score equating methods. The kequate package uses the kernel equating technique supporting Gaussian, logistic, and uniform kernels and performs the IRT observed score equating in the Kernel Equating framework (Andersson, Branberg, & Wiberg, 2013).

The SNSequate package implements the traditional mean, linear, and equipercentile equating methods including the mean–mean, mean–sigma, Haebara, and Stocking–Lord IRT methods. It also supports the newest methods such as local equating, kernel equating, and IRT parameter linking methods based on asymmetric item characteristic functions. Additionally, it also provides the Bayesian non-parametric model for test equating (González, 2014). The irtoys package contains linear equating methods (e.g., mean–mean, mean–sigma, Haebara, and Stocking–Lord methods) to transform a set of IRT parameters (Partchev, Maris, & Hattori, 2017). The TAM package contains the linking method using unidimensional IRT, and linear transformation constants are calculated using Stocking–Lord and Haebara methods (Robitzsch et al., 2018).

Computerized tests

The catR package provides the generation of response patterns under a dichotomous and polytomous computerized adaptive test (CAT) framework (Magis & Raîche, 2012). Item banks also can be generated from prespecified parent distributions for use with dichotomous IRT models. There are options to select the first item(s) and to control for item exposure and content balancing as well as several stopping rules. The CAT can be designed with multidimensional IRT models via the mirtCAT package, which provides tools to generate an HTML interface for creating adaptive and non-adaptive tests using the shiny package (Chalmers, 2016). One can create simple questionnaire forms to collect response data directly in R, and optimal test designs (e.g., shadow testing) are supported. The mstR package contains the procedures to generate response patterns and an item bank under dichotomous and polytomous computerized multistage testing (MST) (Magis, Yan, & von Davier, 2018). The xxIRT package can be used to perform computerized MST, automated test assembly, simulation of CAT, and assembly and simulation of MST (Luo, 2018).

Reliability, classification accuracy, and consistency

The eRm package calculates person separation reliability, and Cronbach's alpha can be calculated by the ltm and Mokken packages. The IRTpp and Mokken packages provide Guttman's lambda, and pairwise package produces WLE person reliability. The Mokken package also estimates various reliability estimates such as Molenaar Sijtsma and latent class reliability coefficient (Heine, 2019; Mair & Hatzinger, 2007; Rizopoulos, 2006; SICS Research Group, 2016; van der Ark, 2007).

When IRT-based scores are used to assess students' achievements, classification accuracy and consistency of assessment play a significant role in key decisions that are made based on grades or classifications (Wheadon, 2014). There are three R packages to calculate classification accuracy and consistency: the cacIRT package using a nonparametric approach, the CDM package using the method of Cui, Gierl, and Chang (2012) and using simulation, and the classify package using IRT approach with bugs or jags replications or plots of score distributions and classification statistics (George et al., 2016; Lathrop, 2015; Wheadon, 2014).

Conclusions

45 R packages using IRT analyses were introduced in this paper, and which types of IRT models and estimation methods can be applied using each were also evaluated. Advanced IRT models such as multidimensional IRT, testlet IRT, mixture IRT, tree-based IRT, hierarchical IRT, IRT based on types of responses, and nonparametric IRT as well as dichotomous and polytomous IRT models can be analyzed via R packages. 34 out of 45 R packages provide simulation study (e.g., data generation, example R codes for simulation, or recovery analysis). These packages make it possible to perform various simulation studies using IRT. In addition, researchers can perform comparison studies on IRT applications (e.g., DIF, equating, and CAT) using different R packages.

ORCID

Youn-Jeng Choi http://orcid.org/0000-0001-9803-2681

References

Andersson, B., Branberg, K., & Wiberg, M. (2013). Performing the Kernel method of test equating with the package kequate. *Journal of Statistical Software*, *55*(6), 1–25. doi:10.18637/jss.v055.i06

Balafas, S., Krijnen, W., Post, W., & Wit, E. (2017). *mudfold: Multiple unidimensional unfolding (R Package version 1.1.1)*. Retrieved from https://cran.r-project.org/package=mudfold

Bartolucci, F., & Bacci, S., (2016). *MLCIRTwithin: Latent class item response theory (LC-IRT) models under within-item multidimensionality (R package version 2.1)*. Retrieved from https://CRAN.R-project.org/package=MLCIRTwithin

Bartolucci, F., Bacci, S., & Gnaldi, M. (2014). MultiLCIRT: An R package for multidimensional latent class item response models. *Computational Statistics & Data Analysis*, *71*, 971–985. doi:10.1016/j.csda.2013.05.018

Battauz, M. (2015). equateIRT: An R package for IRT test equating. *Journal of Statistical Software*, *68*(7), 1–22. doi:10.18637/jss.v068.i07

Brown, A., & Maydeu-Olivares, A. (2013). How IRT can solve problems of ipsative data in forced-choice questionnaires. *Psychological Methods*, *18*(1), 36–52. doi:10.1037/a0030641

Cervantes, V. H. (2017). DFIT: An R Package for Raju's differential functioning of items and tests framework. *Journal of Statistical Software*, *76*(5), 1–24. doi:10.18637/jss.v076.i05

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. doi:10.18637/jss.v048.i06

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–39. doi:10.18637/jss.v071.i05

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2011). lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software*, *39*(8), 1–3.

Cui, Y., Gierl, M. J., & Chang, -H.-H. (2012). Estimating classification consistency and accuracy for cognitive diagnostic assessment. *Journal of Educational Measurement*, *49*, 19–38. doi:10.1111/jedm.2012.49.issue-1

De Boeck, P., & Partchev, I. (2012). IRTrees: Tree-based item response models of the GLMM family. *Journal of Statistical Software*, *48*(1), 1–28. doi:10.18637/jss.v048.c01

Dewey, M. (2018). latdiag: Diagrams for latent trait models. R package version 0.2-4.

Drabinova, A., Martinkova, P., & Zvara, K. (2018). *difNLR: DIF and DDF detection by non-linear regression models (R package version 1.2.2)*. Retrieved from https://cran.r-project.org/web/packages/difNLR/index.html

Fox, J. P., Klotzke, K., & Entink, R. K. (2018). *LNIRT: Log normal response time item response theory models (R package version 0.3.5)*. Retrieved from https://cran.r-project.org/package=LNIRT

Frick, H., Strobl, C., Leisch, F., & Zeileis, A. (2012). Flexible Rasch mixture models with package psychomix. *Journal of Statistical Software*, *48*(7), 1–25. doi:10.18637/jss.v048.i07

George, A. C., Robitzsch, A., Kiefer, T., Groß, J., & Ünlü, A. (2016). The R Package CDM for cognitive diagnosis models. *Journal of Statistical Software*, *74*(2), 1–24. doi:10.18637/jss.v074.i02

González, J. (2014). SNSequate: Standard and nonstandard statistical models and methods for test equating. *Journal of Statistical Software*, *59*(7), 1–30. doi:10.18637/jss.v059.i13

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York, NY: Springer Science+Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Heine, J. H., (2019). *pairwise: Rasch model parameters by pairwise algorithm (R package version 0.4.4-5.1)*. Retrieved from https://cran.r-project.org/web/packages/pairwise

Hohensinn, C. (2018). pcIRT: An R package for polytomous and continuous Rasch models. *Journal of Statistical Software*, *84*(2), 1–14. doi:10.18637/jss.v084.i01

Imai, K., Lo, J., & Olmsted, J. (2017). *emIRT: EM algorithms for estimating item response theory models (R package version 0.0.8*). Retrieved from https://cran.r-project.org/package=emIRT

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2014). Flexible item response theory modeling with FLIRT. *Applied Psychological Measurement*, *38*(5), 404–405. doi:10.1177/0146621614524982

Lathrop, N. Q. (2015). *cacIRT: Classification accuracy and consistency under item response theory (R package version 1.4)*. Retrieved from https://cran.r-project.org/package=cacIRT

Li, Z., & Hong, F. (2014). *plRasch: Log linear by linear association models and Rasch family models by pseudolikelihood estimation (R package version 1.0)*. Retrieved from https://cran.r-project.org/package=plRasch

Luo, X. (2018). *xxIRT: Item response theory and computer-based testing in R (R package version 2.1.1)*. Retrieved from https://cran.r-project.org/package=xxIRT

Magis, D., Beland, S., Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods*, *42*, 847–862. doi:10.3758/BRM.42.3.847

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1–31. doi:10.18637/jss.v048.i08

Magis, D., Yan, D., & von Davier, A. (2018). *mstR: Procedures to generate patterns under multistage testing (R package version 1.2)*. Retrieved from https://cran.r-project.org/web/packages/mstR/mstR.pdf

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), 1–20. doi:10.18637/jss.v020.i09

Mair, P., Hatzinger, R., & Verhelst, N. D. (2015). *RaschSampler: Rasch Sampler (R package version 0.8-8)*. Retrieved from https://cran.r-project.org/package=RaschSampler

Mazza, A., Punzo, A., & McGuire, B. (2014). KernSmoothIRT: An R package for Kernel smoothing in item response theory. *Journal of Statistical Software*, *58*(6), 1–34. doi:10.18637/jss.v058.i06

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin, Germany: De Gruyter.

Mueller, H. (1987). A Rasch model for continuous ratings. *Psychometrika*, *52*, 165–181. doi:10.1007/BF02294232

Oshima, T., Raju, N., & Nanda, A. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement*, *43*(1), 1–17. doi:10.1111/jedm.2006.43.issue-1

Partchev, I., Maris, G., & Hattori, T. (2017). *irtoys: A collection of functions related to item response theory (IRT) (R package version 0.2.1)*. Retrieved from https://cran.r-project.org/package=irtoys

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384. doi:10.3102/10769986027004341

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. *Psychometrika*, *66*(3), 437–459. doi:10.1007/BF02294444

Post, W. J. (1992). *Nonparametric unfolding models: A latent structure approach*. Leiden University, the Netherlands: DSWO Press.

Preinerstorfer, D. (2016). *mRm: An R package for conditional maximum likelihood estimation in mixed Rasch models (R package version 1.1.6)*. Retrieved from https://cran.r-project.org/web/packages/mRm/mRm.pdf

Raiche, G. (2014). *irtProb: Utilities and probability distributions related to multidimensional person item response models (R package version 1.2)*. Retrieved from https://cran.r-project.org/package=irtProb

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proceedings Fourth Berekely Symposium on Mathematical Statistics and Probability*, *5*, 321–333.

Reif, M. (2017). PP: Estimation of person parameters for the 1,2,3,4-PL model and the GPCM. R package version 0.6.1.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), 1–25. doi:10.18637/jss.v017.i05

Robitzsch, A., Kiefer, T., & Wu, M. (2018). *TAM: Test analysis modules. R package version 2.12-18*. Retrieved from https://CRAN.R-project.org/package=TAM

Robitzsch, A., & Steinfeld, J. (2018). *immer: Item response models for multiple ratings. R package version 1.1-35*. Retrieved from https://CRAN.R-project.org/package=immer

Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika*, *38*(2), 203–219. doi:10.1007/BF02291114

Schauberger, G. (2017). *DIFlasso: A penalty approach to differential item functioning in Rasch models*. Retrieved from https://cran.r-project.org/package=DIFlasso

SICS Research Group (2016). *IRTpp: Estimating IRT parameters using the IRT methodology (R package version 0.2.6.1)*. Retrieved from https://cran.r-project.org/package=IRTpp

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.

Sijtsma, K., & Van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology*, *70*, 137–158. doi:10.1111/bmsp.12078

Torres Irribarra, D., & Freund, R. (2014). *Wright Map: IRT item-person map with ConQuest integration*. Retrieved from http://github.com/david-ti/wrightmap

van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19.

Van Schuur, W. H. (1984). *Structure in political beliefs: A new model for stochastic unfolding with application to European party activists*. Amsterdam, the Netherlands: CT Press.

Wheadon, C. (2014). Classification accuracy and consistency under item response theory models using the package classify. *Journal of Statistical Software*, *56*(10), 1–14. doi:10.18637/jss.v056.i10

Willse, J. T. (2014). *mixRasch: Mixture Rasch models with JMLE (R package version 1.1)*. Retrieved from https://CRAN.R-project.org/package=mixRasch

Zes, D., Lewis, J., & Landis, D. (2014). *kcirt: K-cube Thurstonian IRT models (R package version 0.6.0)*. Retrieved from https://CRAN.R-project.org/package=kcirt

Zijlstra, W. P. (2011). *fwdmsa: Forward search for Mokken scale analysis (R package version 0.2)*. Retrieved from https://cran.r-project.org/package=fwdmsa

Zopluoglu, C. (2015). *EstCRM: Calibrating parameters for the Samejima's continuous IRT model (R package version 1.4)*. Retrieved from https://cran.r-project.org/package=EstCRM