Psychological Test and Assessment Modeling, Volume 65, 2023, 259–269

Extending GMX: Conditional Likelihood Ratio Test and Extended Graphical Model Checks with psychotools

Rainer W. Alexandrowicz[1]

Abstract

This article introduces an extension of GMX, which now also supports the conditional likelihood ratio test and graphical model checks for the psychotools package. The package is freely available at `https://osf.io/2ryd8`.

Keywords: Rasch models, graphical model check, conditional maximum likelihood, multi-group split, R-package, psychotools

[1] *Correspondence concerning this article should be addressed to:* rainer.alexandrowicz@aau.at
Rainer W. Alexandrowicz, University of Klagenfurt, Institute of Psychology, Methods Department, Universitaetsstrasse 67–69, 9020 Klagenfurt, Austria.

Introduction

The conditional Likelihood Ratio Test (cLRT; Andersen, 1973) is a statistical test, which allows for checking the adequacy of Item Response Theory (IRT; de Ayala, 2022) models assuming parallel item or threshold characteristic curves, i. e., the Rasch Model (RM; Rasch, 1960), the Partial Credit Model (PCM; Masters, 1982), the Rating Scale Model (RSM; Andrich, 1978) and several descendants of these. In essence, the test compares the equivalence of the parameter estimates of known sub-groups (e. g., low/high score or external split criteria like gender or other groups of substantive interest). Rasch (1960) proposed a Graphical Model Check (GMC) by plotting the parameter estimates of the total sample against those of the sub-samples (e. g., Fig. 2, p. 81; see also Andersen, 1980, p. 257, Fig. 6.2). In current applications, it has become customary to plot the sub-group estimates for two-group splits directly against each other. Such a diagram provides an immediate impression of the similarity of the estimates (optionally complemented by their confidence ellipses) thus supporting the ad hoc identification of problematic items (in the dichotomous case) or thresholds (in the polytomous case).

So far, the only major IRT package in R performing the cLRT and drawing the GMC is `eRm` (Mair, Hatzinger, & Maier, 2020; Mair & Hatzinger, 2007). Its `plotGOF()` routine generates the GMC of the item (RM) or the cumulative threshold (PCM, RSM, LPCM, LRSM) parameter estimates for a two-group split. In a multi-group split, only the parameters of the first two groups are plotted. Therefore, Alexandrowicz (2022) introduced the R package `GMX` (`https://osf.io/2ryd8`), which generates an extended GMC providing the following additional features:

- automated pairwise plots for multi-group splits,
- plotting either the cumulative threshold parameters (the only option of `plotGOF()`), the “standard” threshold parameters, or the person parameters,
- plotting selected split-groups and/or items, and
- several graphic options to flexibly fine-tune the diagram(s), most importantly an automated coloring of all thresholds per item or items per threshold in the polytomous case.

The package takes the return object of `eRm::LRtest()` and has already proven itself a handy tool for easily obtaining informative graphical model checks.

The Extended GMX Package

The new version of GMX also supports the `psychotools` package (Zeileis et al., 2023; Schneider, Strobl, Zeileis, & Debelak, 2022), which like `eRm` is one of the few IRT packages providing Conditional Maximum Likelihood parameter estimation (CML; Baker & Kim, 2004). The `psychotools` package is very flexible and allows for estimating the parameters of many models along with various kinds of graphical output, making it a handy tool for detailed analyses. The CML-based routines for the RM, the PCM, and the RSM are directly implemented, whereas more complex models – like the 2/3/4PL, (Birnbaum, 1968; Barton & Lord, 1981; Lokan & Rulison, 2010), the GPCM, (Muraki, 1992), the GRM, (Samejima, 1969), and many more – are supported via an interface to the `mirt` package (Chalmers, 2012). The package also provides several innovative tools for assessing the model fit. One important option is the anchor based approach (Schneider et al., 2022). However, it currently contains no routine for applying the cLRT or to draw a GMC.

The new function `GMX::cLRT()` performs the conditional likelihood ratio test according to Andersen (1973) for the RM, the PCM, and the RSM using the respective routines of psychotools. The plotting function `GMX::gmx()` has been extended to process the return object of both `eRm::LRtest()` and `GMX::cLRT()`. Figure 1 sketches the workflow:

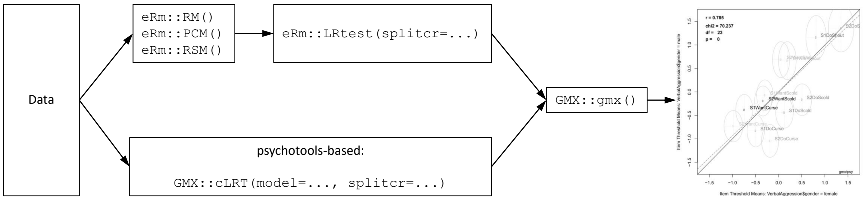

Figure 1: Workflow for building GMCs with GMX.

Moreover, the plotting default has been changed to `type="thresholds"` (rather than `type="betas"`, which was originally chosen for compatibility with `eRm`). To make the results even clearer, the former `type="betas"` option has been renamed to `type="cumulative"` (for the conversion see Equations (1) and (2) in Alexandrowicz, 2022). This renaming was further motivated by the fact that the authors of `psychotools` also use "beta", yet differently: They denote a global difficulty estimate of polytomous items, which is the average of the thresholds per item. To mimic the "`psychotools`-style" betas, `gmx()` now supports a new option `type="means"`, which draws the averaged thresholds per item.

Working Examples

The `cLRT` function requires the data at least. The default model is the PCM and the default split criterion is the median of the score (see Example 1). Possible split criteria are: a vector indicating each row's group membership, a single number used as cut-off for the score, or one of the keywords `"median"` or `"all.r"`, for a median or a full raw score split, respectively.

The examples below use the Verbal Aggression data set of Vansteelandt (2000) (see de Boeck & Wilson, 2004 for a more detailed description, as the original work is not publicly available).

Listing 1: Example using the Verbal Aggression data set of `psychotools`.

```
library(GMX)
> lrt1 = cLRT(VerbalAggression$resp[, 1:12])
[1] "Median split: 9.5"
splitcr
r <= 9.5  r > 9.5      Sum
     158      158      316

The following items have different categories across sub-groups:
 S2DoShout

$S2DoShout
     splitcr
x     r <= 9.5 r > 9.5 Sum
  0        150      88 238
  1          8      45  53
  2          0      25  25
  Sum      158     158 316

The cLRT will be performed with:
 S1WantCurse S1DoCurse S1WantScold S1DoScold S1WantShout S1DoShout
 S2WantCurse S2DoCurse S2WantScold S2DoScold S2WantShout

Andersen cLRT:
    statistic: 27.708
    df:        21
    p-value:   0.1486
```

The `cLRT` routine informs the user that the median of the scores is 9.5 and that the two sub-groups amount to 158 observations each. This example further shows the frequent problem that not all categories appear in all sub-groups. Variable `S2DoShout` has no entries of 2 in the lower score split group and has therefore to be excluded from further analysis. This test yields a non-significant result ($\chi^2 = 27.7$, $df = 21$, $p = .15$).

The unfavorable but necessary omission of the item may be overcome by modifying the cut-off score, because ultimately, we want to compare “lower” to “higher” scores. The choice of the median is just a means to obtain sub-groups of approximately equal size. The `cLRT` routine supports this by giving console feedback so that users immediately know, which cut-off score has not worked out properly. Then, the numeric split option allows for flexibly fine-tuning the cut-off until all items are retained. This has been applied in Example 2. With a cut-off of 11 all categories appear in both split groups. Now, the test reveals a significant result ($\chi^2 = 37.98$, $df = 23$, $p = .03$).

Listing 2: Example applying a user-provided cut-off score.

```
> lrt2 = cLRT(VerbalAggression$resp[, 1:12],11)
[1] "Score split: 11"
splitcr
r <= 11  r > 11     Sum
    196     120     316
Andersen cLRT:
    statistic: 37.984
    df:        23
    p-value:   0.0256
```

The Verbal Aggression data set also contains a gender variable, which may serve as an external split criterion (Example 3).

Listing 3: Example using an external split criterion.

```
> lrt3 = cLRT(VerbalAggression$resp[, 1:12],VerbalAggression$gender)
[1] "Split by  VerbalAggression$gender :"
splitcr
female   male    Sum
   243     73    316
Andersen cLRT:
    statistic: 70.237
    df:        23
    p-value:   0
```

Here, we obtain a significant result indicating gender differences in expression of verbal aggression ($\chi^2 = 70.24$, $df = 23$, $p < .05$). Figure 2 shows an example diagram of how `gmx()` processes the return object of `cLRT()`.

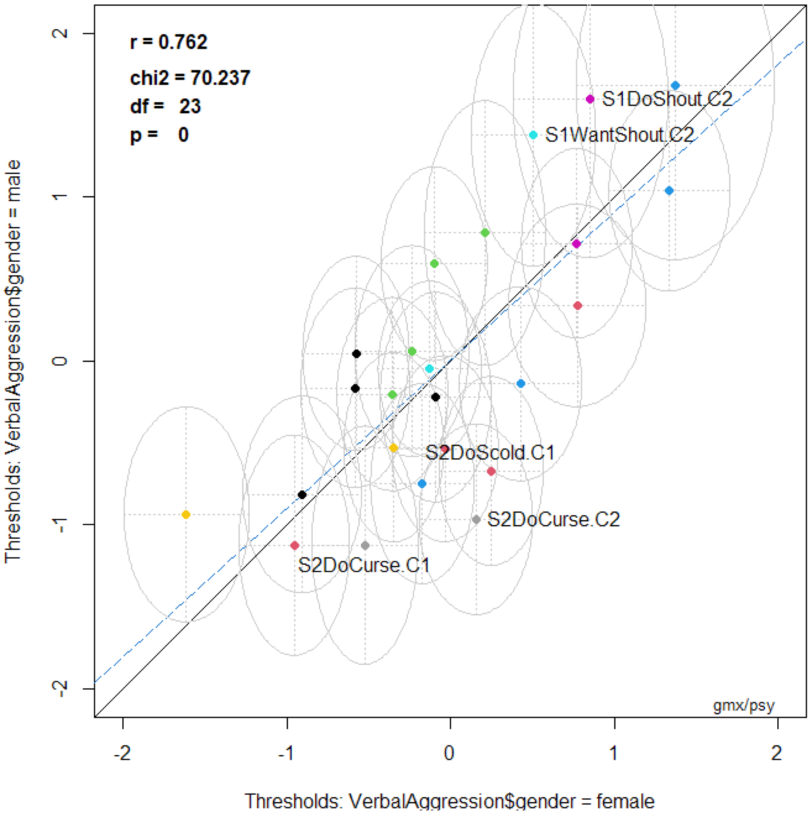

Figure 2: Call: `gmx(lrt3,col="items",tlab="identify")`; Five thresholds have been identified.

The `tlab="identify"` option allows for interactively highlighting thresholds of substantive interest, so that we may point out for example those far away from the identity line. However, the diagram is still somewhat cluttered. Here, the new option `type="means"` comes in handy by showing just one point per item (Figure 3).

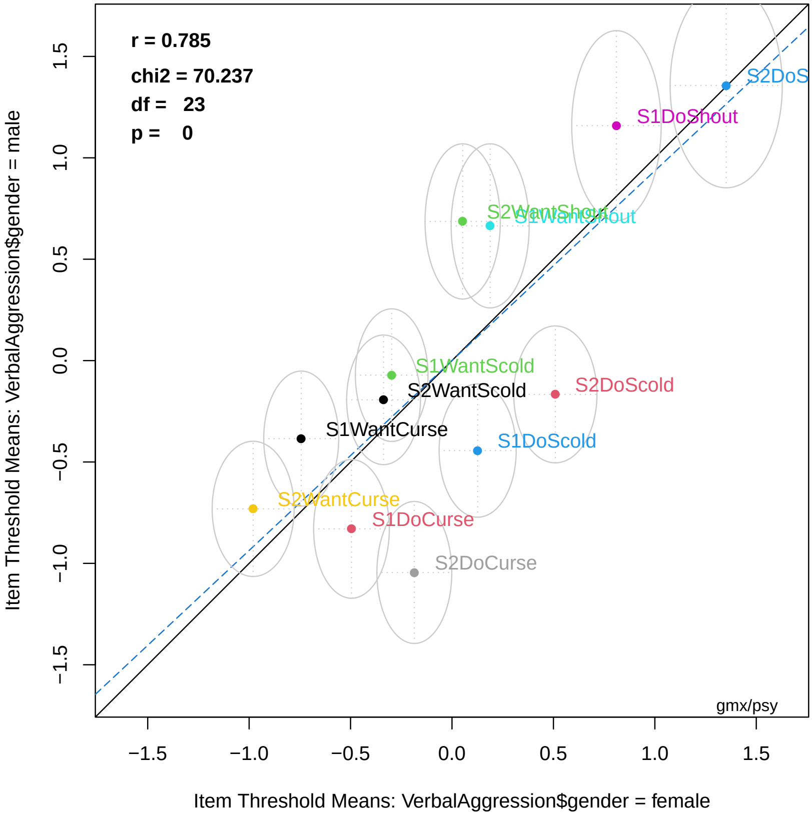

Figure 3: Call: `gmx(lrt3,type="means",col="items",tlab="names")`

We now see that, for example, cursing is easier for men compared to women, whereas wanting to shout shows the opposite pattern. This result is in line with de Boeck and Wilson (2004, p. 375). Further graphical options of `GMX` are extensively discussed in Alexandrowicz (2022).

Discussion

The extended GMX package now supports the conditional likelihood ratio test using psychotools routines by providing the function GMX::clrt(). Its output object is processed by the drawing function GMX::gmx() so that all features of the GMX package can be utilized by users preferring psychotools over eRm. Thus, GMX blends seamlessly into the various further options of model tests and checks provided by the psychotools package.

The decision to make type="thresholds" the new default was made because the eRm default may give rise to misunderstanding. Users may not be aware that the "betas" of eRm denote the cumulative thresholds. In fact, it seems hard to imagine a situation in which the cumulative thresholds would be advantageous. Accordingly, the axis labels have been changed from previously "Betas" to "Cumulative Thresholds" to inform users better about what is actually drawn.

A particular problem arises for estimating the parameters of a PCM with CML, when categories of an item have never been used. The psychotools package offers three options to handle such null categories via the nullcats option, which accepts the arguments "keep", "downcode", and "ignore". The first (and default) applies the procedure described in Wilson and Masters (1969), the second fills the gap by downcoding the available categories above the missing one(s), and the third excludes these items from the analysis. As has become apparent in Example 1 may the omission of items change the result drastically. As the eRm package only supports the item omission option, differences between the two packages occur depending on the chosen option. The cLRT() function supports the "..." argument, which allows for passing further options to the estimation routines. Thus, the flexibility of the psychotools package remains applicable.

Another difference between the two supported packages is the specific way the latent scale is identified. This results in different diagrams for polytomous items if the cumulative threshold option is chosen. However, no attempt was made to equalize the diagrams, because each program offers various identification options[1]. Moreover, plotting the cumulative thresholds seems of little practical interest for the reasons mentioned above. Nevertheless, the cLRT results are the same except for differences in handling null categories. For the same reason, also the person parameter estimates for polytomous responses may differ. Again, assimilation is futile, because the set of person parameters only make sense in relation to the set of item parameters, whichever standardization

[1] In the course of testing GMX, a bug in the eRm::RSM() function has become apparent. See: https://r-forge.r-project.org/tracker/index.php?func=detail&aid=6805&group_id=80ß&atid=363.

has been chosen. Finally, users will likely prefer either the `eRm` or the `psychotools` package, but rarely switch.

A cautionary note seems indicated: The polytomous models (i. e., PCM and RSM) require zero-based integer coding (i. e., 0,1,…, $m_j - 1$, with m_j denoting the number of categories of item j). Both packages detect the frequently applied coding $1, 2, \ldots, m_j$ and automatically shift the codes downwards (with a message). However, this automatism may turn out counterproductive, if the shift is not applied to all split-groups. This may easily happen if the zeroes appear only in the other split-group(s). Therefore, it is advisable to check for the correct coding in advance rather than relying on "smart" software.

The plot annotation default option has also been changed to `annot=c("r","LRT")` in this new version, now adding both the correlation coefficient of the plotted parameters and the results of the cLRT (When multiple plots are drawn, the cLRT results will only be added to the first one). In the former version, only the correlation was added by default. Users will thus have all relevant information available at a glance. The empty vector `annot=c("")` or `annot=c("none")` suppresses any annotation.

Especially the `color="item"` and `color="thresholds"` feature of `gmx()` have proven useful, as they allow for quickly highlighting all thresholds of each item or all first, second, …thresholds, respectively. The `eRm::plotGOF()` also supports that kind of highlighting, but one had to apply handicrafts to build the according color vector by counting parameters.

This new version of GMX provides users of both major R packages supporting the CML estimation method with easy to use routines to perform graphical model checks in various ways.

Acknowledgements The author thanks Achim Zeileis and Rudolf Debelak from the `psychotools` developer team for quickly fixing minor bugs, which have become apparent during the making of `GMX`. Version 0.7-3 or later of `psychotools` should be used. The author thanks further Milica Kabic and Davide Albers for beta testing the new package version.

References

Alexandrowicz, R. W. (2022). GMX: Extended Graphical Model Checks A Versatile Replacement of the plotGOF() Function of eRm. *Psychological Test and Assessment Modeling*, *64*, 215–225.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Andersen, E. B. (1980). *Discrete Statistical Models with Social Science Applications*. North-Holland.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory. Parameter Estimation Techniques.* NY: Marcel Dekker.

Barton, M. A., & Lord, F. M. (1981). *An Upper Asymptote for the Three-Parameter Logistic Item-Response Model.* Educational Testing Service Research Report Series [RR-81-20]. doi: 10.1002/j.2333-8504.1981.tb01255.x

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statsitical Software*, *48*(6), 1–29.

de Ayala, R. J. (2022). *The Theory and Practice of Item Response Theory* (2nd ed.). The Guilford Press.

de Boeck, P., & Wilson, M. (2004). *Explanatory Item Response Models. A Generalized Linear and Nonlinear Approach.* Springer.

Lokan, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *British Journal of Mathematical and Statistical Psychology*, *63*, 509–525. doi: 10.1348/000711009X474502

Mair, P., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science*, *49*, 26–43.

Mair, P., Hatzinger, R., & Maier, M. J. (2020). eRm: Extended Rasch Modeling [Computer software manual]. Retrieved from https://cran.r-project.org/package=eRm (1.0-2)

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*, 149–174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Journal of Educational Measurement*, *16*(2), 159–176.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danmarks Pædagogiske Institut.

Samejima, F. (1969). Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometrika Monograph Supplement*.

Schneider, L., Strobl, C., Zeileis, A., & Debelak, R. (2022). An R toolbox for score-based measurement invariance tests in IRT models. *Behavior Research Methods*, *54*, 2101–2113. doi: 10.3758/s13428-021-01689-0

Vansteelandt, K. (2000). *Formal models for contextualized personality psychology* (Unpublished doctoral dissertation). K.U.Leuven, Belgium.

Wilson, M., & Masters, G. N. (1969). The Partial Credit Model and Null Categories. *Psychometrika*, *58*, 87–99.

Zeileis, A., Strobl, C., Wickelmaier, F., Komboz, B., Kopf, J., Schneider, L., & Debelak, R. (2023). psychotools: Infrastructure for psychometric modeling [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=psychotools (R package version 0.7-3)