Psychological Test and Assessment Modeling, 2025.67:166-175
DOI: https://doi.org/10.2440/001-0023

On the bias of Andersen's conditional Likelihood Ratio test for the Rasch model given data sampled according to adaptive testing[1]

Klaus D. Kubinger

University of Vienna, Faculty of Psychology

Abstract:
A key feature of the Rasch model is that it enables specific objective comparisons of the (psychological) test items' difficulties after the items have been administered to examinees. At the end, that property allows a means of testing the model. Based on the respective empirical data, the hypothesis can be statistically tested and hence either accepted or rejected that the model holds for the given pool of items. The most widely used statistical test in this context is Andersen's (conditional) Likelihood Ratio test, which relies on conditional maximum likelihood (CML) item parameter estimation – there the condition applies that the still unknown ability parameters of the examinees are to substitute by the respective sufficient statistic, i.e. the raw score, the number of solved items. However, already Glas (1988) revealed that CML item parameter estimation yields biased results when items are administered through "multi-stage" testing, or to say, when "branched" (adaptive) testing applies. Although Kubinger, Steinfeld, Reif, and Yanagida (2012) showed that the ability parameter estimations' corresponding percentile ranks most widely coincide with those of the true ability parameters (that is, the item parameter estimation bias, due to CML estimation, does not have a relevant effect for case consulting), the respective effect on Andersen's Likelihood Ratio test has not yet been demonstratively illustrated. The question at hand is whether its test statistic would then actually hold the type-I-risk. The result of the presented simulation study is as follows: Rasch model-conforming data obtained *via* branched testing produces an actual distribution of Andersen's Likelihood ratio test-statistic with a 99-quantile which is up to ten times (and even) larger than that in the suggested χ^2-distribution. Consequently, model misfit would be very often falsely

Correspondence: Klaus D. Kubinger, c/o University of Vienna, Faculty of Psychology, Liebiggasse 5, A-1010-Vienna, Austria, klaus.kubinger@univie.ac.at, www.klaus-kubinger.com

[1] Acknowledgement: This paper was conceptualized from the very beginning together with *Jan Steinfeld*, who finally did all the statistical analyses and simulation studies.

concluded. And the same phenomenon, although only half so extensive occurred when Rasch model-conforming data were obtained *via* tailored testing. While in the case of branched adaptive testing the problem solution is to use for the CML parameter estimation the now available R-program *tmt: Estimation of the Rasch model for multistage tests* (Steinfeld & Robitzsch, 2019), this program does not work in the case of tailored adaptive testing: in the latter case, only the here given approach is on hand, that is to proceed a targeted simulation study.

Keywords: Rasch model, tailored testing, Likelihood Ratio test, conditional maximum likelihood (CML) estimation, multi-stage testing

Introduction

The importance of the (dichotomous) Rasch model (Rasch, 1960/1980; see also Fischer, 1974) for psychological test calibration is of no doubt nowadays. Nevertheless, it is always worthwhile to remind of the mathematical law: "If the number of solved items should be a fair measure of the examinees' ability [i.e. should be a (minimal) sufficient statistic] then the Rasch model must hold" (for a proof, see Fischer, 1974, 1995). It is also well known that the Rasch model allows specific objective comparisons of items (as well as examinees; see in particular Scheiblechner, 2009), implying measurements that put objects, e.g. examinees, in empirically adequate relationships to one another without taking into account which other objects have been or will be considered. That feature of specific objective measurement allows also a means of testing the model. Based on the respective empirical data, the hypothesis can be statistically tested and hence either accepted or rejected that the model holds for the given item pool. This goes beyond pertinent goodness-of-fit indices, which only indicate the extent to which the data can be explained by the model. That is, the (absolute) validness of the model is concerned and not only the (relative) goodness-of-fit either in comparison to other competing models, or how well the model is able to describe the data. There are several approaches of testing the Rasch model based on this concept of "specific objectivity" (for a good overview, see Glas & Verhelst, 1995). Although there are two additional approaches dealing even with determining the necessary sample size when the type-I- and the type-II-risk as well as the relevant effect of model contradiction are given (see on the one hand Draxler, 2010, Draxler & Alexandrowicz, 2015, and Alexandrowicz & Draxler, 2016, and on the other hand Kubinger, Rasch, & Yanagida, 2009, 2011, and Yanagida, Kubinger, & Rasch, 2015 – see for a collectively overview Draxler & Kubinger, 2018), the most commonly applied test in this context is Andersen's (conditional) Likelihood Ratio test (Andersen, 1973). This test opposes the data's likelihood when item parameter estimation is based on the whole sample with that one when item parameter estimation is based on some partitioned subsamples; given "specific objectivity", both these likelihoods do not significantly differ. We will deal with this test in the following.

The background of our considerations goes back to Glas (1988). He already revealed that pertinent Rasch model item calibration yields biased item parameter estimations

when items are administered using "multi-stage" testing. "Pertinent" in this case means that in order to utilize the Rasch model's feature of allowing specific objective comparisons of the items as the basis for testing model validness, conditional maximum likelihood (CML) estimation is needed. Multi-stage testing or to say "branched" (adaptive) testing means that every examinee is only administered step by step with those groups (blocks) of few items each, the item parameters of which best fit with the current ability parameter estimation. Hence, if CML estimation is indispensable for the given reason, the claim by Eggen and Verhelst (2011) to use in the case of branched testing marginal maximum likelihood (MML) estimation instead for item parameter estimation is of no assistance. Admittedly, Kubinger, Steinfeld, Reif, and Yanagida (2012) relativized the problem a bit. Using biased item difficulty parameter estimations, their simulation study shows: while ability parameter estimations when branched testing is applied are similarly biased, the estimations' corresponding percentile ranks of the simulees most widely coincide with those of the true ability parameters. In other words, the item parameter estimation bias (due to CML estimation), which occurs when estimates are based on item administration according to some branched testing design, does not result in relevant effects for case consulting. That is, examinees will be assigned almost exactly with the same percentile rank whether biased or unbiased ability parameter estimations were used; i.e. the relative position of all the examinees within the population in question remains the same.

However, there is no illustrative evidence of how Andersen's Likelihood Ratio test reacts on biased item parameter estimations. The matter is whether its test-statistic, which is supposed to be asymptotically χ^2-distributed, actually holds the type-I-risk when the item parameter estimations are biased due to the fact that the data were sampled by branched testing[2] – in particular applying the sample's partition criterion low vs. high score, which commonly leads to the most powerful model test.

The biased CML item parameter estimation with branched testing arises as follows: Each examinee is administered with a specific combination of (groups of) items due to his/her current test achievement and finally gains a certain (raw) score, i.e. the number of solved items. But independent of the resulted item combination and the score, there can not occur every theoretically possible response pattern (of solved and not solved items) – otherwise, the branching from (groups of) items to (group of) items would have contradicted the current test achievement of the examinee (Kubinger, Steinfeld, Reif, & Yanagida, 2012, give an elaborated illustration of this fact, exemplarily). However, calculation of the data's likelihood (within the CML item parameter estimation process) refers foundamentally to the score-conditioned probability of the observed response pattern, being based on all theoretically possible response patterns and standardized/weighted accordingly (cf. the so-called elementary symmetric function; Rasch, 1060/1980, or Fischer, 1981). That is, the formula for the likelihood

[2] Zwitser & Maris (2015) only remarks in their paper, that the "LRT [Andersen's Likelihood Ratio test] may then become conservative" (p. 74).

to maximize is just a misspecification if some of the theoretically possible response patterns have principally a probability of zero.[3]

Fortunately, to some extent the basic problem has been psychometrically solved in the meantime: Zwitser and Maris (2015) deduced the formula of the score-conditioned probability of an observed response pattern, given a (very) special design of branching due to the examinee's preceding test achievement. And they even provide researchers with a computer program that appropriately estimates the item parameters for specified multi-stage/branched testing designs. Furthermore, while this program does not calculate Andersen's Likelihood Ratio test the R-program *tmt: Estimation of the Rasch model for multistage tests* (Steinfeld & Robitzsch, 2019) does.

With this respect, the problem of parameter estimation has basically already been solved (Steinfeld & Robitzsch, 2024). Nevertheless, it seems necessary to point this problem out because psychological test constructors may not be aware of it and, for instance using the excellent and world-wide applied R-program *eRm* (*extended Rasch modeling*; Mair, Rusch, Hatzinger, Maier, & Debelak, 2025) for their Rasch model-based test calibration, may believe in the test-statistic's (asymptotically claimed) χ^2-distribution of Andersen's Likelihood Ratio test.[4]

That is, in this paper we analyze this test-statistic's actual type-I-risk, given biased CML item parameter estimations due to branched testing. We will do this by means of a simulation study. And we also consider the case of the data being sampled with tailored testing, i.e. obviously a generalization of multi-stage testing as there even after each item that one is administered next which item parameter best fits with the current ability parameter estimation (cf. e.g. Kubinger, 2016). Though there is a much greater number of sub-pools of administered items, the issue quoted above nevertheless arises, that is, given a certain item combination, actually not every response pattern can occur.

[3] Bear in mind that these troubles do not correspond with the case using so-called test-booklets, where an (very large) item pool is partitioned into several non-disjunctive item sub-pools, each of which is administered to a different (randomly allocated) group of examinees – given such test-booklets have been conceptualized according to a connected incomplete block design: of course unbiased CML estimation results, as was first established in the FORTRAN-program LLTM (Formann, 1974) and perpetuated in all its succeeding programs, particularly in eRm (extended Rasch modeling; Mair & Hatzinger, 2007, Mair, Hatzinger, & Maier, 2015, Mair, Rusch, Hatzinger, Maier, & Debelak, 2025).

[4] Several earlier manuscript drafts of this paper were submitted elsewhere years ago. Due to a very lengthy review process, the cited computer programs got published in the meantime. And also the approach given below to simulate the actual density of the Andersen's Likelihood Ratio test statistic in the discussed case of adaptive testing has in the meantime been applied for a relevant test calibration (see Kubinger & Holocher-Ertl, 2014; Kubinger, 2017; Kubinger, 2023).

Method

We focused on a very specific branched testing design (see Fig. 1), which is not only a rather complex one and therefore particularly illustrative, but is also applied widely in practice (see e.g. Kubinger, 2017). 60 items are arranged according to some item difficulty levels into non-disjunctive sub-pools of five items each (blocks). There are three stages, that is, every examinee is administered three times five items. The starting point is due to some ability level grouping, based on prior information (age of the examinee). That is, there is not a general sub-pool of items for starting the first stage, but actually four such pools. 3-way branching applies: if the examinee has responded correctly to a maximum of one of the respective five items, then he/she is administered with an easier item sub-pool next; if he/she has responded correctly to at least four items, then a more difficult sub-pool is presented next; and if he/she responds correctly to two or three items (roughly half of the items), then a sub-pool is administered which matches the (mean) item difficulty parameter of the preceding item sub-pool.

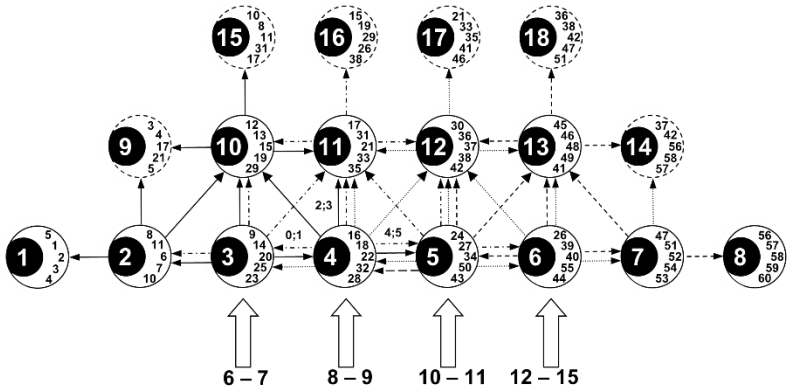


Figure 1: The branched testing design that is being considered. The circles represent different item sub-pools (blocks), arranged from left to right according to their average item difficulty parameter. Each item sub-pool contains five items; the code number of each corresponds to the rank order of all the items' difficulty parameters. Based on the examinee's age, the first stage starts with another sub-pool. The next respective item sub-pool is then chosen according to the examinee's performance (that is, according to the number of items responded correctly to in the preceding item sub-pool). Given the examinee is eight or nine years old, the Figure shows in detail which combinations of item sub-pools could occur, depending on previous performances. Branched testing terminates after three stages, that is after the third item sub-pool; for this reason, the dashed-lined circles also consist of items from the solid-lined ones. (From Kubinger, 2017; with kind permission of Hogrefe Publishing).

The simulation study was planned and conducted as follows. We used the empirical given but almost uniform distributed item difficulty parameters in the interval from

-9.35 to 6.09 (similar to Kubinger, Steinfeld, Reif, & Yanagida, 2012, but now in accordance with Kubinger, 2023). The ability parameters were supposed to be mixed-distributed with respect to the ability level groups (age; i.e. 6 to 15), that is normally distributed with the same standard deviation of 1.5 but with different means: -3.17, -2.55, -1.18, -0.23, 0.73, 1.59, 2.50, 2.87, 3.58, and 3.77 (this due to the empirical data as cited above). The simulation of Rasch model conforming data according to branched testing was then done by using the R-package *eRm* (see above). The sample size of simulees was fixed with 10,000 for a first scenario and with 4,000 for a second one. For both scenarios 3,500 runs were done. Each resulting data set was analyzed with *eRm*, as well. That is, the biased CML item parameter estimation was applied for the Andersen's Likelihood Ratio test calculation – using the partition criterion low vs. high score. At the end, we counted the number of runs out of the 3,500, which resulted in significance for the 99-quantile of the χ^2-distribution (i.e., the nominal type-I-risk $\alpha = .01$). The relative frequency of this case estimates the actual type-I-risk.

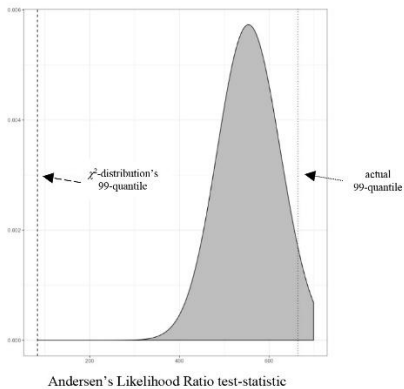
For some control, we did a corresponding simulation study, but now all simulees were administered with every item; that is, a conventional scenario with fixed item testing. As then CML item parameter estimation is unbiased, the test-statistic of Andersen's Likelihood Ratio test should actually be (asymptotically) χ^2-distributed. Finally, we performed an analogue simulation study when tailored testing was applied. Starting with the item having a medium difficulty parameter with respect to the first stage's group-according sub-pool, those items which best fit the actual ability parameter estimation of the simulee were selected step by step (we used the approach suggested by Warm, 1989, at the very beginning; basically, the R-package *PP* of Reif & Steinfeld, 2017, was applied). Tailored testing terminated when two consecutive parameter estimations did not differ more than 0.7 or when 15 items had already been administered. Again, we counted the number of runs out of the 3,500, which resulted in significance for the 99-quantile of the χ^2-distribution, in order to estimate the actual type-I-risk of Andersen's Likelihood Ratio test. All simulations were done in R (R Core Team, 2016).

Results

Table 1 compares the simulation results of all the three different test principles, that is branched, fixed, and tailored testing, with respect to: a) the 99-quantile of Andersen's Likelihood Ratio test-statistic's actual distribution, and b) the accordingly actual type-I-risk of this test. Additionally, the 99-quantile of the χ^2-distribution is given, which theoretically applies. As not all of the 3,500 runs deliver data that lead to parameter estimations for all items, the number of degrees of freedom of Andersen's Likelihood Ratio test varies slightly. Table 1 always refers to the number of items and degrees of freedom, respectively, which resulted as a maximum. For the case of the test principles of branched and tailored testing Figure 2 presents in addition the actual distribution graphically, using a scenario with 1,000 runs and 4,000 simulees.

Table 1: The 99-quantile of Andersen's Likelihood Ratio test-statistic's actual distribution for branched, fixed, and tailored testing, when CML item parameter estimation applies. "Actual" means: based on a simulation study (3,500 runs). The 99th-quantile of the claimed χ^2-distribution is also given. [In brackets, the number of runs (r) with the respective degree of freedom (df).]

	branched testing			fixed testing			tailored testing		
	actual type-I-risk	actual 99-quantile	χ^2-distribution's 99-quantile	actual type-I-risk	actual 99-quantile	χ^2-distribution's 99-quantile	actual type-I-risk	actual 99-quantile	χ^2-distribution's 99-quantile
10,000 simulees	1.00	1511.05	84.73 ($r = 1226$; $df = 57$)	.01	79.55	81.07 ($r = 1069$; $df = 54$)	1.00	817.47	87.17 ($r = 2878$; $df = 59$)
4,000 simulees	1.00	664.22	83.51 ($r = 1344$; $df = 56$)	.01	78.28	78.62 ($r = 1173$; $df = 52$)	1.00	364.37	85.95 ($r = 3442$; $df = 58$)



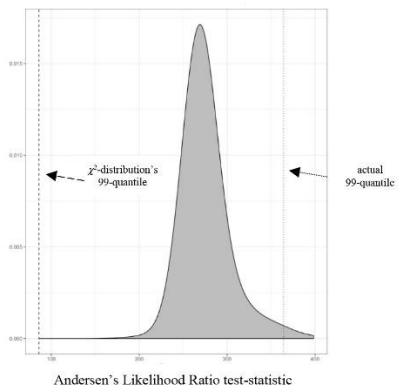


Figure 2: The actual distribution (the 99-quantile included) of Andersen's Likelihood Ratio test-statistic for branched testing (left side) and for tailored testing (right side), when CML item parameter estimation applies (1,000 runs and 4,000 simulees). The 99-quantiles of the claimed χ^2-distribution are also shown.

The main result is that, given data sampled according to branched testing, the CML item parameter estimation bias must not at all be underestimated. In contrast with Kubinger, Steinfeld, Reif, and Yanagida (2012), who established rather minor consequences concerning the percentile ranks of examinees – which are exclusively of practical relevance – severe errors do indeed arise when deciding whether an item pool is or is not in accordance with the Rasch model: while Rasch model conforming data, based on administration according to the branched testing design of Figure 1, discloses

an actual distribution of Andersen's Likelihood ratio test-statistic with a 99-quantile of 664.22 (4,000 simulees for each of 3,500 runs), the respective critical value of the χ^2-distribution ($df = 56$) only amounts to 83.51. As a consequence of this, the use of the χ^2-distribution produces a lot of artificial significance.

As assumed, fixed testing does not cause any problems. Andersen's Likelihood ratio test-statistic holds the type-I-risk when the χ^2-distribution is applied.

As far as tailored testing is concerned, also an enormous bias of the distribution of Andersen's Likelihood Ratio test-statistic occurs, though the 99-quantile of 364.37 amounts rather only a half of branched testing one's. The bias is due to the fact that some response patterns for several sub-pools of administered items realistically also have a probability of zero.

Discussion

Although we exemplified the problem of artificial results of Andersen's Likelihood ratio test when data were sampled by branched testing only in a special but very complex design, there is no doubt that for every design, the χ^2-distribution does not at all meet the actual distribution of the respective test-statistic. Ignoring this circumstance means that in many cases a model violation would artificially result. Admittedly, this phenomenon is widely known within the psychometricians' community, as mentioned here accordingly – though hardly within the community of test authors and test publishers. However, it seems rather unknown that the same phenomenon occurs if data were used, sampled by tailored testing.

To summarize, the problem of artificial results of Andersen's Likelihood ratio test when data were sampled by branched testing, can be easily overcome by using the R-program *tmt*. However, when the data were sampled by tailored testing only the approach used here can be applied, i.e., the simulation of the respective 99- (or 95-) quantile according to some postulated parameters for a Rasch model conforming item pool. The reason is that the Zwitser and Maris (2015) approximation of bias correction will then not work because the information for each relevant sub-pool of administered items is far too small.

References

Alexandrowicz, R.W., & Draxler, C. (2016). Testing the Rasch model with the conditional likelihood ratio test: sample size requirements and bootstrap algorithms. Journal of Statistical Distributions and Applications*, 3*, article no. 2 (2015). https://doi.org/10.1186/s40488-016-0039-y.

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

Draxler C. (2010). Sample size determination for Rasch model tests. *Psychometrika, 75,* 708-724.

Draxler, C., & Alexandrowicz, R.W. (2016). Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. *Psychometrika, 80*, 897-919. doi: 10.1007/s11336-015-9472-y.

Draxler, C., & Kubinger, K.D. (2018). Power and sample size considerations in psychometrics. In J. Pilz, D. Rasch, V. B. Melas, & K. Moder (eds.), *Statistics and Simulation* (pp. 39-51). Heidelberg: Springer.

Eggen, T.J.H.M., & Verhelst, N.D. (2011). Item calibration in incomplete testing designs. *Psicológica, 32*, 107-132.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* [Introduction into theory of psychological tests]. Berne: Huber.

Fischer, G.H. (1981). On the existence and uniqueness of maximum likelihood estimates in the Rasch model. *Psychometrika, 46,* 59-77.

Formann, A.K. (1974). Programm zur Schätzung der Basisparameter des dichotomen logistischen Testmodells: LLTM [Computer program for the estimation of the elementary operation parameters of the dichotomous logistic test model: LLTM]. In G.H. Fischer, *Einführung in die Theorie psychologischer Tests* [Introduction into theory of psychological tests] (pp. 531-554). Berne: Huber.

Glas, C.A.W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics, 13*, 45-52.

Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch Model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models* (pp. 69-95). New York: Springer.

Kubinger, K.D. (2016). Adaptive testing. In K. Schweizer & C. DiStefano (eds.), *Principles and methods of test construction* (pp. 104-119). Göttingen: Hogrefe.

Kubinger, K.D., (2017). *Adaptive Intelligence Diagnosticum 3 – English Edition (AID 3)*. Göttingen & Oxford: Hogrefe.

Kubinger, K.D., (2023). *Manual und Testhandbuch zum AID 3 (Version 3.2) von K. D. Kubinger & S. Holocher-Benetka* [Manual and Instruction of the AID 3, version 3.2, by K. D. Kubinger & S. Holocher-Benetka]. Göttingen: Hogrefe.

Kubinger, K.D., & Holocher-Ertl, S. (2014). *Adaptives Intelligenz Diagnostikum - Version 3.1 (AID 3)* [Adaptive Intelligence Diagnosticum – version 3.1 (AID 3)]. Göttingen: Beltz.

Kubinger, K.D., Rasch, D., & Yanagida, T. (2009). On designing data-sampling for Rasch model calibrating an achievement test. *Psychology Science Quarterly, 51,* 370-384.

Kubinger, K.D., Rasch, D., & Yanagida, T. (2011). A new approach for testing the Rasch model. *Educational Research and Evaluation, 17,* 321-333.

Kubinger, K.D., Steinfeld, J., Reif, M., & Yanagida, T. (2012). Biased (conditional) parameter estimation of a Rasch model calibrated item pool administered according to a branched-testing design. *Psychological Test and Assessment Modeling, 54,* 450-461.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: the eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*, 1-20.

Mair, P., Hatzinger, R., & Maier, M.J. (2015). eRm: Extended Rasch modeling. R package version 0.15-5.

Mair, P., Rusch, T., Hatzinger, R., Maier, M.J., & Debelak, R. (2025). eRm: Extended Rasch modeling. R package version 1.0-9.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Reif, M., & Steinfeld, J. (2017). PP: Estimation of person parameters for the 1,2,3,4-PL model and the GPCM. R package version 0.6.1. https://github.com/manuelreif/PP.

Steinfeld, J., & Robitzsch, A. (2019). tmt: Estimation of the Rasch model for multistage tests (R Package Version 0.2.1-0). https://CRAN.R-project.org/package=tmt

Steinfeld, J., & Robitzsch, A. (2024). Conditional maximum-likelihood estimation in probability-based multistage designs. *Behaviormetrika, 51*, 617–634. https://doi.org/10.1007/s41237-024-00228-3

Scheiblechner, H.H. (2009). Rasch and pseudo-Rasch models: suitableness for practical test applications. *Psychology Science Quarterly, 51*, 181-194.

Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427-450.

Yanagida, T., Kubinger, K.D., & Rasch, D. (2015). Planning a study for testing the Rasch model given missing values due to the use of test-booklets. *Journal of Applied Measurement*, *16*, 432-444.

Zwitser, R.J., & Maris, G. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika*, *80*, 65-84.