Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 19

4-18-2019

The Andersen Likelihood Ratio Test with a Random Split Criterion Lacks Power

Georg Krammer
University College of Teacher Education Styria, georg.krammer@phst.at

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2685
doi: 10.22237/jmasm/1555594442

ISSN 1538 − 9472

The Andersen Likelihood Ratio Test with a Random Split Criterion Lacks Power

Georg Krammer
University College of Teacher Education Styria
Graz, Austria

The Andersen LRT uses sample characteristics as split criteria to evaluate Rasch model fit, or theory driven hypothesis testing for a test. The power and Type I error of a random split criterion was evaluated with a simulation study. Results consistently show a random split criterion lacks power.

Keywords: Andersen LRT, Rasch model, split criteria, power, Type I error

Introduction

By means of the Andersen likelihood ratio test (LRT; Andersen, 1973), the person homogeneity of the Rasch model (Rasch, 1960) can be assessed. The LRT uses split criteria based on test scores (e.g., median split), or external criteria (e.g., gender). Any split criteria can be used, and it was suggested that the type of split criteria affects the power of the LRT (Glas & Verhelst, 1995; Gustafsson, 1980; Rost, 1990; van den Wollenberg, 1979). Such a split criterion may even be a random split (Hambleton & Murray, 1983). Moreover, choosing a split criterion that is essentially meaningless would also constitute a random split (Molenaar, 1983); this would be the case if, for example, gender was used repeatedly as a split criterion without a theoretical basis to why person homogeneity could be violated across genders.

Simulation studies so far have been limited to the LRT with a median split (e.g. Alexandrowicz & Draxler, 2016; Futschek, 2014; Gustafsson, 1980; Suárez-Falcón & Glas, 2003). The present simulation study addresses this gap in research. The aim of the study is to shed light on Type I error and power of the LRT when a median split is not used, but a random split criterion.

doi: 10.22237/jmasm/1555594442 | Accepted: December 23, 2017; Published: April 18, 2019.
Correspondence: Georg Krammer, georg.krammer@phst.at

Split Criteria

The most commonly-used split criterion for the LRT is the median split. Scholars have argued that an appropriate split criterion should be related to performance, i.e., test takers' raw scores (e.g., Andersen, 1982; Andrich, 1978; Glas & Verhelst, 1995). However, the split criterion does not have to be based on the test takers' scores, and using only score based split criteria may mask model misfit (e.g., Gustafsson, 1980; Rost, 1990; van den Wollenberg, 1982).

Consequently, the LRT has been used with a multitude of external split criteria, either to test global item fit of the Rasch model, or for theory-driven hypothesis testing. For example, it was used to assess person homogeneity across commonly-used external split criteria such as age and gender; across various test properties: response format (Hohensinn & Kubinger, 2011), language (Arendasy, Sommer, & Mayr, 2012), test-taking time (Gittler & Fischer, 2011), and item order (Ortner, 2004); across educational variables: educational degree (van de Grift, Helms-Lorenz, & Maulana, 2014), and length of schooling (Schultz-Larsen, Kreiner, & Lomholt, 2007); across nationalities and languages (Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011; Kreiner & Christensen, 2014; Lauritsen, Kreiner, Söderström, Dørup, & Lous, 2015; Yang et al., 2011); across health related issues and physiological criteria: previous strokes and types of housing (Schultz-Larsen et al., 2007) and middle ear status (Lauritsen et al., 2015); and even across workplace conditions such as school type and pupils in classrooms (van de Grift et al., 2014). A theoretical basis is not always given for why person homogeneity across given subsamples is being tested.

The LRT has also been used with a random division of given samples, i.e., a random split criterion. Such a split into random subsamples was first proposed as part of a graphical inspection of the invariance of 1-PL Rasch model item parameters (Hambleton & Murray, 1983), and was soon employed for the LRT (e.g., Maier & Philipp, 1985, 1986; Maier, Philipp, Buller, & Schiegel, 1987). There are numerous examples of using a random split (e.g., Devy, Lehert, Varlan, Genty, & Edan, 2015; Gnambs & Batinic, 2011; Kliem et al., 2015; Koller & Alexandrowicz, 2010; Rusch, Mair, Lowry, & Treiblmaier, 2013). However, little is known about Type I error and power of an LRT with a random split. So far, only tentative evidence has been offered that an LRT with a random split has less power in detecting multidimensionality as compared to the median split (Schoppek & Landgraf, 2011). No evidence has been offered regarding violations of the parallel ICC assumption or the local independence assumption.

The Current Study

In summary, the LRT is used with a multitude of external split criteria among a random division of the sample. However, Type I error and power of the LRT with any split criteria other than the median have never been systematically addressed. Therefore, the aim of the current simulation study is to scrutinize the use of the random split for the LRT. An LRT with a random split is expected to have less power than with a median split (e.g., Schoppek & Landgraf, 2011). The results are expected to shed light on using the random split in general, but also on using essentially meaningless external split criteria, i.e. split criteria for which no theoretical basis for their use as split criteria is given.

Methodology

Data were simulated under 105 conditions. In line with the most exhaustive simulation study addressing the LRT (*cf*. Suárez-Falcón & Glas, 2003), data were simulated adhering to the 1-PL Rasch model and data violating assumptions of the 1-PL Rasch model. Three types of 1-PL Rasch model violations (no parallel ICCs, no local independence, or no unidimensionality) were simulated. Data were simulated with different test lengths (10, 25, and 50 items) and sample sizes (100, 250, 500, 1000, and 1500). Additionally, two degrees of 1-PL Rasch model violation were simulated: high and moderate (*cf*. Suárez-Falcón & Glas, 2003) for the three types of model violations.

To violate the parallel ICC assumption, data were simulated according to a 2-PL model (*cf*. Birnbaum, 1968). The discrimination parameters of each item were drawn from a lognormal distribution ($M = 0$) with a standard deviation of 0.5 or 0.25, corresponding to a high and a moderate degree of model violation, respectively. To violate the local independence assumption, a pairwise inter-item correlation was simulated. The pair-wise inter-item correlation was either 1 or .5 for all consecutive pairs of items, corresponding to a high and a moderate degree of model violation, respectively. To violate the unidimensionality assumption, two-dimensional Rasch model data were simulated. For this two-dimensional data, the correlation between the two factors was either 0 or .5, corresponding to a high and a moderate degree of model violation, respectively.

For each condition, 1000 data sets were simulated using the Extended Rasch Modeling package (eRm; Mair & Hatzinger, 2007). For each simulated data set, an LRT with a random split was computed. The random split was based on the random number generation of R: a random vector was used to assign every person either to the first or second subsample. As benchmark for comparison, an LRT with median split was also computed. The test statistic of the LRT was computed on the basis of the conditional maximum-likelihood of the whole sample and of the two subsamples (*cf*. Andersen, 1973), and evaluated against a .05 significance level.

Results

Shown in Table 1 are the absolute number of significant ($p < .05$) LRT for each condition and type of split criteria. The significant LRT in the first column (1-PL Rasch model data) represent the Type I error; in the other columns (1-PL Rasch model violations), they represent power. Although the results were comparable for the Type I error rates across the types of split criteria, clear differences can be seen in the power analysis.

Type I Error

The Type I error of the LRT (the 1-PL column of Table 1) did not differ from the nominal-level (50 out of 1000) for the median split (the upper half of Table 1: $M = 49.4$, $SD = 9.2$, t[14] = −0.25, $p = .80$) and the random split (the lower half of Table 1: $M = 52.7$, $SD = 9.1$, t[14] = 1.14, $p = .28$). Thus, the LRT discards as many data fitting the 1-PL Rasch model as it should, irrespectively of the type of split criterion.

Power

For the LRT with a random split, there was no discernible pattern in the change of power depending on the type of model violation, the degree of model violation, the test length, or even the sample size. Moreover, the power was non-existent in every condition, in the best cases only fairly exceeding the nominal level. In contrast, the power analysis for the LRT with a median split was as expected: the power was higher the larger the sample size, the longer the test length, and the higher the degree of model violation. In line with previous simulation studies, the power of the LRT was the highest in detecting violations of the parallel ICC assumption (*cf*. Suárez-Falcón & Glas, 2003). In summary, the LRT with a median split performed as expected, while the LRT with a random split did very poorly in comparison: the power of an LRT with random splits more closely resembled a Type I error than sensitivity against model violations.

Table 1. Number of significant (p < .05) LRT for each condition; the LRT were computed with a split at the median (Median) and a random split (Random); 1-PL Rasch model assumptions (1-PL), 2-PL model assumptions (2-PL), local dependencies (Loc. Dep.), and two-dimensionality (2-dim) were simulated for different sample sizes (n), test lengths (k), and degrees of model violation

Split criterion	k	n	1-PL	2-PL ln(0, 0.5)	2-PL ln(0, 0.25)	Loc. dep. δ = 1	Loc. dep. δ = .5	2-dim r = 0	2-dim r = .5
Median	10	100	46	337	138	58	44	167	70
		250	45	743	269	61	54	380	100
		500	44	929	573	109	54	521	185
		1000	51	989	843	205	76	643	343
		1500	51	998	923	289	130	741	426
	25	100	63	406	289	70	52	215	90
		250	49	1000	721	123	70	345	117
		500	49	1000	969	201	86	493	220
		1000	65	1000	1000	380	111	614	325
		1500	37	1000	1000	595	189	738	416
	50	100	61	990	522	89	68	194	84
		250	55	1000	966	156	84	309	109
		500	51	1000	1000	300	110	481	196
		1000	42	1000	1000	623	157	598	287
		1500	32	1000	1000	839	268	662	375
Random	10	100	69	53	66	59	45	67	62
		250	44	35	53	60	63	48	55
		500	48	49	59	62	50	52	58
		1000	51	55	43	56	51	51	54
		1500	59	61	52	55	49	68	54
	25	100	55	66	56	53	65	55	61
		250	41	62	55	52	62	61	58
		500	49	48	53	57	63	55	55
		1000	45	56	48	55	56	63	48
		1500	50	62	40	44	51	57	52
	50	100	69	61	46	63	52	71	84
		250	64	46	40	66	70	86	54
		500	52	45	50	45	52	67	63
		1000	41	50	45	54	57	75	62
		1500	53	65	62	51	55	86	55

Note: ln(M, SD) = lognormal distribution with mean M = 0 and $SD \in \{0.5, 0.25\}$; δ = pair-wise inter-item correlation; r = factor correlation; 1000 data sets were simulated for each condition

Conclusion

The results demonstrated consistently for all types of model violations, samples sizes, and test lengths that an LRT with a random split lacks power. Researchers are well advised not to utilize the LRT with a random split. On a cautionary note, any other split criteria than the median should be well-grounded in theory. If meaningless split criteria are chosen, the LRT will nearly always accept the person homogeneity of the compared subsamples.

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*(2). doi: 10.1186/s40488-016-0039-y

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*(1), 123-140. doi: 10.1007/bf02291180

Andersen, E. B. (1982). Latent trait models and ability parameter estimation. *Applied Psychological Measurement, 6*(4), 445-461. doi: 10.1177/014662168200600406

Arendasy, M. E., Sommer, M., & Mayr, F. (2012). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Journal of Cross-Cultural Psychology, 43*(3), 464-479. doi: 10.1177/0022022110397360

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.

Devy, R., Lehert, P., Varlan, E., Genty, M., & Edan, G. (2015). Improving the quality of life of multiple sclerosis patients through coping strategies in routine medical practice. *Neurological Sciences, 36*(1), 85-90. doi: 10.1007/s10072-014-1900-8

Futschek, K. (2014). Actual type-I- and type-II-risk of four different model tests of the Rasch model. *Psychological Test and Assessment Modeling, 56*(2), 168-177.

Gittler, G., & Fischer, G. (2011). IRT-based measurement of short-term changes of ability, with an application to assessing the "Mozart Effect". *Journal of Educational and Behavioral Statistics, 36*(1), 33-75. doi: 10.3102/1076998610366260

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent* *developments, and applications* (pp. 69-95). New York, NY: Springer. doi: 10.1007/978-1-4612-4230-7_5

Gnambs, T., & Batinic, B. (2011). Evaluation of measurement precision with Rasch-type models: The case of the short Generalized Opinion Leadership Scale. *Personality and Individual Differences, 50*(1), 53-58. doi: 10.1016/j.paid.2010.08.021

Gustafsson, J. E. (1980). Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 33*(2), 205-233. doi: 10.1111/j.2044-8317.1980.tb00609.x

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory* (pp. 71-94). Vancouver, BC: Educational Research Institute of British Columbia.

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732-746. doi: 10.1177/0013164410390032

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*(6), 497-509. doi: 10.1080/13803611.2011.632668

Kliem, S., Beller, J., Kröger, C., Stöbel-Richter, Y., Hahlweg, K., & Brähler, E. (2015). A Rasch re-analysis of the Partnership Questionnaire. *SAGE Open, 5*(2). doi: 10.1177/2158244015588958

Koller, I., & Alexandrowicz, R. W. (2010). Eine psychometrische analyse der ZAREKI-R mittels Rasch-modellen [A psychometric analysis of the ZAREKI-R using Rasch models]. *Diagnostica, 56*(2), 57-67. doi: 10.1026/0012-1924/a000003

Kreiner, S., & Christensen, K. B. (2014). Analyses of model fit and robustness. A new look at the PISA scaling model underlying ranking of countries according to reading literacy. *Psychometrika, 79*(2), 210-231. doi: 10.1007/s11336-013-9347-z

Lauritsen, M. B. G., Kreiner, S., Söderström, M., Dørup, J., & Lous, J. (2015). A speech reception in noise test for preschool children (the Galker-test): Validity, reliability and acceptance. *International Journal of Pediatric Otorhinolaryngology, 79*(1), 1694-1701. doi: 10.1016/j.ijporl.2015.07.028

Maier, W., & Philipp, M. (1985). Comparative analysis of observer depression scales. *Acta Psychiatrica Scandinavica, 72*(3), 239-245. doi: 10.1111/j.1600-0447.1985.tb02601.x

Maier, W., & Philipp, M. (1986). A polydiagnostic scale for dimensional classification of endogenous depression derivation and validation. *Acta Psychiatrica Scandinavica, 74*(2), 152-160. doi: 10.1111/j.1600-0447.1986.tb10599.x

Maier, W., Philipp, M., Buller, R., & Schiegel, S. (1987). Reliability and validity of the Newcastle Scales in relation to ICD-9-classification. *Acta Psychiatrica Scandinavica, 76*(6), 619-627. doi: 10.1111/j.1600-0447.1987.tb02932.x

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). doi: 10.18637/jss.v020.i09

Molenaar, I. W. (1983). Some improved diagnostics for failure of the Rasch model. *Psychometrika, 48*(1), 49-72. doi: 10.1007/bf02314676

Ortner, T. M. (2004). On changing the position of items in personality questionnaires: Analysing effects of item sequence using IRT. *Psychology Science, 46*(4), 466-476.

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

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271-282. doi: 10.1177/014662169001400305

Rusch, T., Mair, P., Lowry, P. B., & Treiblmaier, H (2013, December). *Developing and measuring IS scales using item response theory*. Proceedings of the 2013 International Conference on Information Systems, Milan, Italy, 15-18 December (pp. 1-16.).

Schoppek, W., & Landgraf, A. (2011). Can a multidimensional hierarchy of skills generate data conforming to the Rasch model? A comparison of methods. *Psychological Test and Assessment Modeling, 53*(1), 3-34.

Schultz-Larsen, K., Kreiner, S., & Lomholt, R. K. (2007). Mini-Mental Status Examination: Mixed Rasch model item analysis derived two different cognitive dimensions of the MMSE. *Journal of Clinical Epidemiology, 60*(3), 268-279. doi: 10.1016/j.jclinepi.2006.06.007

Suárez-Falcón, J. C., & Glas, C. A.W. (2003). Evaluation of global testing procedures for item fit to the Rasch model. *British Journal of Mathematical and Statistical Psychology, 56*(1), 127-143. doi: 10.1348/000711003321645395

van de Grift, W., Helms-Lorenz, M., & Maulana, R. (2014). Teaching skills of student teachers: Calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*, 150-159. doi: 10.1016/j.stueduc.2014.09.003

van den Wollenberg, A. L. (1979). *The Rasch model and time-limit tests: An application and some theoretical contributions* (Unpublished doctoral dissertation). Nijmegen, Netherlands: Katholieke Universiteit te Nijmegen. Retrieved from https://repository.ubn.ru.nl/handle/2066/147865

van den Wollenberg, A. L. (1982). A simple and effective method to test the dimensionality axiom of the Rasch model. Applied Psychological Measurement, 6(1), 83-91. doi: 10.1177/014662168200600109

Yang, F. M., Heslin, K. C., Mehta, K. M., Yang, C. W., Ocepek-Welikson, K., Kleinman, M., ... & Jones, R. N. (2011). A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos. *Psychological Test and Assessment Modeling, 53*(4), 440-460.