Psychological Test and Assessment Modeling, Volume 58, 2016 (1), 3-13

Editorial
Review and forecast on research in Psychological Test and Assessment Modeling

Klaus D. Kubinger (editor in chief)[1]

Preamble: The last editorial (cf. Kubinger, 2014) emphasized the great attention which this journal, focusing on "Psychological Test and Assessment Modeling" since 2010, receives from researchers interested in psychology-specific statistical methods and problems, general psychometrics, and psychological assessment in theory and practice. Now, two years later, we can indeed try to become indexed by Thomson-Reuters due to our very encouraging, though fluctuating self-evaluated impact factor (according to Kubinger, Heuberger, & Poinstingl, 2010); 2010: 0.565, 2011: 0.525, 2012: 0.783, 2013: 0.420, 2014: 0.354, 2015: 0.370. In order to expand the distinguishing character of this journal, we now request the authors to commit themselves to making the following available on demand: a) the data, b) the specification of the used software (version number and applied options included), as well as c) the applied source code if no pertinent software is used – this is due to the standards of research reproducibility established by Hothorn and Leisch (2011).[2]

Introduction

In the following editorial, we once again outline the scope of the journal, but primarily give indications on how to manage research work in order to contribute to the concerning area at a very high methodical standard. We deal with three topics once more: i) *statistical standards*, which have been partially raised in the meantime and which may concern future research by statisticians or psychologists and others with proper expertise in this

[1] *Correspondence concerning this article should be addressed to:* Prof. Dr. Klaus D. Kubinger, c/o Division for Assessment and Applied Psychometrics, Faculty of Psychology, University of Vienna, Liebiggasse 5, A-1010 Vienna, Austria; email: klaus.kubinger@univie.ac.at

[2] We have slightly changed the editorial board: in doing so Markus Bühner, Edgar Erdfelder, Alexander von Eye, and Manfred Prenzel left the board. We kindly thank them for all their competent support in editing the journal and for contributing to the image of it. On the other hand, we are very pleased to welcome: David Andrich, Alison Ying Chen, Paul De Boeck, Andreas Frey, and Samuel Greiff.

area. ii) *psychometric standards*, which experience new insights almost daily, particularly due to large scale assessment research and the option of simulation studies. iii) *psychological assessment proceedings*, concerning either new approaches of modeling and measuring traits or dealing with new (psycho-) technological elaborations.

For each topic, such papers published in this journal in the last two years (i.e. 2014 and 2015) are quoted which contribute essentially to psychological test and assessment modeling. Furthermore, an outlook seems appropriate on which effort is to be expected in research work.

Standards and proceedings

Statistical standards

In general, none of the already stated misuses of statistical analyses and improper respective traditions within psychology have been abandoned yet (cf. Rasch, Kubinger, and Yanagida, 2011 – if the reader prefers German, see Kubinger, Rasch, and Yanagida, 2011). That is, above all a) the "practice of asterisks", which always implies the highest α of all α-levels one would ever accept – if a researcher tries to impose the result's conclusiveness on the reader by this means then matter-of-factly to quote the estimated effect size is only informative; b) furthermore, the arbitrary choice of the type-I-risk (α) without reflecting the consequences of an eventual type-II-error – instead, calculating *a-posterior* the "result-based type-II-risk" at least approaches the state-of-the-art, which is to calculate the sample size: hereby only relevant effects will result in significance, but such relevant effects will not be detected with the probability of some settled type-II-risk only (most favorable use the R-routine OPDOE [*OPtimal Design Of Experiments*], Rasch, Pilz, Verdooren, & Gebhardt, 2011); c) the ignorance toward sequential testing, which is the preferred method if the data are sampled one after the other, because it generally saves quite a lot of sample size (the R-routine OPDOE serves for such analyses as well); d) pre-testing the theory-based assumptions of normal distributions and homogeneity of the variances when applying the two-sample *t*-test or the analysis of variance, though these tests lead to unknown final type-I- and type-II-risks if performed using the same set of observations (which is usually the case) – instead it is recommended to apply the Welch-test as a standard test and Hotelling's T^2, respectively; and finally e) the lack of insight, that a significant correlation coefficient is hardly of any use as even a correlation coefficient of .01 can reach significance, given the sample size is large enough – instead, only the determination coefficient is of any meaning (i.e. the effect size in question) with the consequence of this being better to calculate the sample size in advance based on a certain type-I- and type-II-risk and a value of the determination coefficient which is of practical relevance (at least, the null-hypothesis H_0: $0 < \rho \leq \rho_0$, e.g. $\rho_0 = .70$, rather than the null-hypothesis H_0: $\rho = 0$ should be tested).

Concerning the latter, Schneider, Rasch, Kubinger, and Yanagida (2015) established a sequential (triangular) test of a correlation coefficient's null-hypothesis H_0: $0 < \rho \leq \rho_0$.

The respective computer program for its application will be integrated into the R-routine OPDOE soon.

Admittedly, planning a study, given a certain type-I- and type-II-risk as well as a relevant effect size, is currently only at a researcher's disposal for parametric tests and almost only for univariate analyses. While planning according to a parametric test although a non-parametric homologous test is targeted seems reasonable, concerning multivariate analyses in general, we nowadays have hardly more than the suggestion of Rasch, Kubinger, and Yanagida (2011, p. 418): "Planning a study according to a multivariate analysis of variance happens either with regard to an in some way 'most important' [variable]; or the researcher calculates the necessary sample size for each [variable] on its own – given certain precision requirements – and then decides for the largest one. However, neither type-I- nor type-II-risk will be kept with regard to the research as a whole (i.e. research-wise risk)." Nevertheless, particularly the discriminant analysis is an example for a more satisfying situation: Regarding the maximum error of predicted assignment to one of the groups according to the resulting discriminant function, the necessary sample size can be calculated in order to fulfill any given type-I-, type-II-risk or effect size (see Rasch, Herrendörfer, Bock, Victor, & Guiard, 2008) – though there still is no computer program for its application. Therefore, quite a lot of research work seems to be needed; but see the respective approaches for testing the Rasch model below (i.e. *Psychometric Standards*).

Admittedly, sequential testing is only at a researcher's disposal for parametric tests. And it is still not elaborated for multivariate analyses; again, there is currently only the suggestion to apply this approach with regard to an in some way "most important" variable.

On the other hand, we can report an important contribution in statistics regarding linear structure equation models (LSEM). Themessl-Huber (2014; in this journal) investigated the appropriateness of the pertinent χ^2-statistic as well as of several fit-indices, with respect to confirmatory factor analysis, by a simulation study. He proved that the χ^2-statistic does fairly fail the type-I-risk; and only the cut-off values provided by *Hu & Bentler* of the CFI (comparative fit index) are somewhat adequate.

Warne and Larsen (2014) found in their simulation study, that for ascertaining the number of factors in exploratory factor analysis, *Guttman*'s traditional rule (number of eigenvalues larger than 1) is less accurate not only in comparison to *Velicer*'s minimum average partial approach and *Horn*'s parallel analysis, but also to their own approach; this extends the traditional rule in regarding rather the eigenvalues' confidence intervals instead of the pure eigenvalues: the confidence interval must exceed 1 in order to indicate a relevant factor.

Psychometric standards

Research at and based on Item Response Theory (IRT) is booming more than ever; this is particularly due to large-scale assessments' establishment in the society. We can divide respective research work and applications as follows: either we focus on the Rasch model's property of specific objective comparisons – especially with regard to the estimation

of the item (difficulty) parameters; or we consider all the IRT-models just as special cases of the general linear model. From the point of philosophy of science, the former approach seems superior as ultimately it ensures means of *testing a model* going 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 in relation to other, competing models. On the other hand, referring to the general linear model allows, of course, almost unlimited extensions for any relevant item parameters as well as generalizations to multidimensional measuring items – even modeling some correlation parameters is within that frame. Contributions have been published to both backgrounds in the last two years in this journal.

Heine and Tarnai (2015) reactivated a pairwise (conditional) item parameter estimation for the Rasch model due to the practical reality of missing values (by chance); their simulation study proved that even for a high rate of missing values, this approach leads to proper item parameter estimations. Futschek (2014) investigated several Rasch model tests according to their type-I- and type-II-risk; apart from already partially known results, the simulation study provides precise information of the quite high power of the Martin-Löf test, given model contradiction due to multidimensionality. Finch and French (2014) did a DIF (differential item functioning) analysis by means of a simulation study with regard to *Birnbaum*'s guessing parameter in the 3-PL model; given a respective effect, they established a severe parameter estimation bias with respect to both, item difficulty and item discrimination parameter. In her simulation study, DeMars (2015) examined the detectability of DIF if at least two groups do not differ by a constant item difficulty shift, but this shift is also a random variable with a certain standard deviation; using *Mantel-Haenszel* DIF procedure for both cases, the results showed hardly any differences. Hagquist and Andrich (2015) investigated the phenomenon of artificial DIF, which is an item favoring one group may induce the appearance of a non-existing DIF in another item favoring another group; the result demonstrates that the magnitude of such an artifact depends on the respective item difficulty parameters in relation to the distribution of the person ability parameters. Salzberger (2015) suggested a test investigating whether the thresholds in polytomous Rasch models are to be considered truly ordered or disordered, based on standard errors of threshold estimations. Kröhne, Goldhammer, and Partchev (2014) performed a simulation study to compare two approaches with respect to ability parameter estimations' efficiency in multidimensional adaptive testing; when item administration is constrained to a pre-specified order of dimensions (one after the other instead of intermixing items from different dimensions), this approach is generally not disadvantageous. Ranger and Kuhn (2014) suggested a goodness-of-fit index for jointly modeling responses and response times; according to their simulation study, it holds the type-I-risk and has high power. George and Robitzsch (2014) introduced an adapted estimation routine for cognitive diagnosis models (see von Davier, 2010, in this journal), as these models often become (nearly) non-identifiable with a growing number of modeled sub-competencies. Irribarra, Diakow, Freund, and Wilson (2015) generalized the approach of Formann (1995), who specified Latent Class analysis in such a way that within each latent group the items measure unidimensional according to the Rasch model; they use *Masters*' Partial Credit model instead of the Rasch model. Wind (2015)

reminded, in accordance with Kubinger (2005), that data might not fit the Rasch model because they substantially follow the deterministic Guttman-scale, the respective extent of conformity quantifiable by the *Mokken* analysis' holomorphy index; she now illustrates the application for polytomous data. Vidotto, Vermunt, and Kaptein (2015) explained the general logic for the use of Latent Class analysis for an imputation, given missing categorical data in Large Scale assessment studies, and illustrated the practical application. And Vink, Lazendic, and van Buuren (2015) showed how to partition the data for traditional imputation techniques in order to manage large data. Rose, von Davier, and Nagengast (2015) dealt with non-ignorable item non-responses, that is omitted or not reached items in psychological or educational tests; they derived respective multidimensional IRT models. Glas, Pimentel, and Lamers (2015) generalized the approach of adequately taking non-ignorable item non-responses into account for polytomous data and even incorporate covariates for the appearance of missing in their model; a simulation study illustrates the efficiency of the model. Aßmann, Gaasch, Pohl, and Carstensen (2015) dealt with missing values in background variables within Large Scale assessment studies, for which they apply a Bayesian estimation strategy using the conditional distribution of the missing values; their simulation study evaluated the respective appropriateness.

Although not in this journal, Yanagida, Kubinger, and Rasch (2015) dealt with the determination of sample size according to a given type-I- and type-II-risk and a certain effect of model contradiction when testing the Rasch model for the case of using several test-booklets, that is there are missing values by design. Their approach for complete data, published in the predecessor of this journal (Kubinger, Rasch, & Yanagida, 2009), proved to work then as well.

This topic of sample size determination within IRT analyses seems to set a trend (cf. Draxler, 2010, as well as Draxler & Alexandrowicz, 2015). Furthermore, investigations of various test-statistics applied within psychometrics seem particularly worthwhile, in order to see whether they actually hold the type-I-risk.

Psychological assessment proceedings

Admittedly, proceedings in psychological assessment based on modeling the interdependencies of personal traits and context variables were rather rare in the last two years – in general as well as in this journal. Using LSEM, Greiff, Krkovic, and Nagy (2014) examined whether two postulated task characteristics explain the mastering of complex problems; this comes close to the LLTM (linear logistic test model)-tradition (Fischer, 1973, 2005; see also Kubinger, 2008, in the predecessor of this journal), which hypothesizes some elementary cognitive operations being specifically responsible for an item's difficulty. Similarly, using a linear regression model, Aryadoust (2015) tried to explain Rasch model parameters for listening comprehension items by some postulated metacognitive strategies.

Obviously, pertinent LLTM-analyses would mean a proper means of testing any model of interest within psychological assessment.

As concerns new (psycho-) technological elaborations, the Decision-Oriented interview by Westhoff (2014), as well as by Westhoff and Hagemeister (2014), seems to be a prototype, published in this journal; the authors suggest a hierarchy of topics and questions in oral examinations, to examine whether given requirements are fulfilled by a candidate – empirical results prove objective and valid assessments according to this technique. Zhou and Reckase (2014) elaborated optimal designing an item pool for adaptive testing; using the Partial Credit model extended with a discrimination item parameter, they simulated theoretical item pools even with practical constraints of content balancing and item exposure control. And Khorramdel (2014) completed an experiment in order to oppose two different rating scale formats; within high stakes assessment, results proved less faking tendencies in personality questionnaires for 6-point instead of 2-point rating scales.

Nevertheless, any effort to improve psychological instruments' validity and accuracy of measurement, the economy and reasonableness of their administration, and their fairness particularly regarding intercultural and globalized effects, seems of importance.

Several papers have been published in this journal dealing with concrete instruments for psychological assessment. Schweizer and Reiß (2014) analyzed a neuroticism scale's validity by means of LSEM; Baghaei and Grotjahn (2014) tried to establish the construct validity of an (English) conversational C-Test by applying the Rasch model; Vladut, Vialle, and Ziegler (2015), as well as Paz-Baruch (2015), validated a questionnaire of educational and learning resources and Leana-Taşcılar (2015) analyzed the same questionnaire due to sex and age differences; Etzler, Rohrmann, and Brandt (2014) investigated a published anger inventory with respect to its validity for inmates; Galić, Scherer, and LeBreton (2014) proved culture-based DIFs for an objective personality test of aggression; similarly French, Hand, Nam, Yen, and Vazquez (2014) found culture-based DIFs with respect to a critical thinking test; Proyer, Wagner-Menghin, and Grafinger (2014) developed a reading comprehension test and Yanagida, Strohmeier, Toda, and Spiel (2014) introduced a questionnaire of individualism/collectivism; Reutlinger, Ballmann, Vialle, Zhang, and Ziegler (2015) offered a questionnaire tracking down the expectations and goals of parents of kindergarten children; Harder, Trottler, Vialle, and Ziegler (2015) presented a teacher and a parent checklist for assessing a student's resources for learning.

Concerning methods for validation, Schweizer (2014) illustrated how to use confirmatory factor analyses for the multitrait-multimethod approach.

Finally, some papers dealt with the typical behavior of testees during psychological assessment. Geiser, Okun, and Grano (2014) investigated volunteer motivation; Hotulainen, Thuneberg, Hautamäki, and Vainikainen (2014) examined how attention measured in prolonged over-learned response tasks correlates with reasoning and school achievement; Tirp, Steingröver, Wattie, Baker, and Schorer (2015) investigated the transferability of specific skills between virtual and real learning environments; Steinbach and Stöger (2015) analyzed the influence of parent's attitudes toward self-regulated learning and the actual achievement behavior; while Niederkofler, Herrmann, Seiler, and Gerlach (2015) investigated influences of students' perception of class climate on

achievement motives, Steuer and Dresel (2015) investigated the influence of classrooms' error climate on achievements; Bollmann, Böbel, Heene, and Bühner (2015) established the inter-individually differentiating variable of benefiting from the true *vs.* false response format in achievement, instead of the pertinent multiple-choice response format; and last but not least Köhler, Pohl, and Carstensen (2015) demonstrated that persons' missing propensities may be regarded as person-specific.

All these efforts continue to be of interest.

Postscript: Authors are warmly encouraged also to publish new computer routines (particularly done in R) which support Psychological Test and Assessment Modeling.

References

Aßmann, C., Gaasch, C., Pohl, S., & Carstensen, C.H. (2015). Bayesian estimation in IRT models with missing values in background variables. *Psychological Test and Assessment Modeling, 57,* 595-618.

Aryadoust, V. (2015). Application of evolutionary algorithm-based symbolic regression to language assessment: Toward nonlinear modeling. *Psychological Test and Assessment Modeling, 57,* 301-337.

Baghaei, P., & Grotjahn, R. (2014). Establishing the construct validity of conversational C-Tests using a multidimensional Rasch model. *Psychological Test and Assessment Modeling, 56,* 60-82.

Bollmann, S., Böbel, E., Heene, M., & Bühner, M. (2015). Which person variables predict how people benefit from True-False over Constructed Response items? *Psychological Test and Assessment Modeling, 57,* 147-161.

DeMars, C.E. (2015). Modeling DIF for simulations: Continuous or categorical secondary trait? *Psychological Test and Assessment Modeling, 57,* 279-300.

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

Draxler, C., & Alexandrowicz, R.W. (2015). Sample size determination within the scope of conditional maximum likelihood estimation with special focus on testing the Rasch model. *Psychometrika, 80,* 897-919.

Etzler, S.L., Rohrmann, S., & Brandt, H. (2014). Validation of the STAXI-2: A study with prison inmates. *Psychological Test and Assessment Modeling, 56,* 178-194.

French, B.F., Hand, B., Nam, J., Yen, H.J., & Vazquez, J.A.V. (2014). Detection of Differential Item Functioning in the Cornell Critical Thinking Test across Korean and North American students. *Psychological Test and Assessment Modeling, 56,* 275-286.

Galić, Z., Scherer, K.T., & LeBreton, J.M. (2014). Examining the measurement equivalence of the Conditional Reasoning Test for Aggression across U.S. and Croatian samples. *Psychological Test and Assessment Modeling, 56,* 195-216.

Geiser, C., Okun, M.A., & Grano, C. (2014). Who is motivated to volunteer? A latent profile analysis linking volunteer motivation to frequency of volunteering. *Psychological Test and Assessment Modeling, 56,* 3-24.

George, A.C., & Robitzsch, A. (2014). Multiple group cognitive diagnosis models, with an emphasis on differential item functioning. *Psychological Test and Assessment Modeling, 56,* 405-432.

Glas, C.A.W., Pimentel, J.L., & Lamers, S.M.A. (2015). Nonignorable data in IRT models: Polytomous responses and response propensity models with covariates. *Psychological Test and Assessment Modeling, 57,* 523-541.

Greiff, S., Krkovic, K., & Nagy, G. (2014). The systematic variation of task characteristics facilitates the understanding of task difficulty: A cognitive diagnostic modeling approach to complex problem solving. *Psychological Test and Assessment Modeling, 56,* 83-103.

Finch,W.H., & French, B.F. (2014). The impact of group pseudo-guessing parameter differences on the detection of uniform and nonuniform DIF. *Psychological Test and Assessment Modeling, 56,* 25-44.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G.H. (2005). Linear logistic test models. In *Encyclopedia of Social Measurement, 2,* 505-514.

Formann, A. K. (1995). Linear logistic latent class analysis and the Rasch model. In G.H. Fischer, & I. Molenaar (eds), *Rasch models: Foundations, recent developments, and applications* (pp. 239–255). New York: Springer.

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,* 168-177.

Hagquist, C., & Andrich, D. (2015). Determinants of artificial DIF – a study based on simulated polytomous data. *Psychological Test and Assessment Modeling, 57,* 342-376.

Harder, B., Trottler, S., Vialle, W., & Ziegler, A. (2015). Diagnosing resources for effective learning via teacher and parent checklists. *Psychological Test and Assessment Modeling, 57,* 201-221.

Heine, J.H., & Tarnai, C. (2015). Pairwise Rasch model item parameter recovery under sparse data conditions. *Psychological Test and Assessment Modeling, 57,* 3-36.

Hothorn, T., & Leisch, F. (2011). Case studies in reproducibility. *Briefings in Bioinformatics, 12*, 288-300.

Hotulainen, R., Thuneberg, H., Hautamäki, J., & Vainikainen, M.P. (2014). Measured attention in prolonged over-learned response tasks and its correlation to high level scientific reasoning and school achievement. *Psychological Test and Assessment Modeling, 56,* 237-254.

Irribarra, D.T., Diakow, R., Freund, R., & Wilson, M. (2015). Modeling for directly setting theory-based performance levels. *Psychological Test and Assessment Modeling, 57,* 396-422.

Khorramdel, L. (2014). The influence of different rating scales on impression management in high stakes assessment. *Psychological Test and Assessment Modeling, 56,* 154-167.

Köhler, C., Pohl, S., & Carstensen, C.H. (2015). Investigating mechanisms for missing responses in competence tests. *Psychological Test and Assessment Modeling, 57,* 499-522.

Kröhne, U., Goldhammer, F., & Partchev, I. (2014). Constrained Multidimensional Adaptive Testing without intermixing items from different dimensions. *Psychological Test and Assessment Modeling, 56,* 348-367.

Kubinger, K.D. (2005). Psychological Test Calibration using the Rasch Model - Some Critical Suggestions on Traditional Approaches. *International Journal of Testing, 5,* 377-394.

Kubinger, K.D. (2008). On the revival of the Rasch model-based LLTM: From constructing tests using item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50,* 311-327.

Kubinger, K. D. (2014). Editorial: Toward essential contributions for Psychological Test and Assessment Modeling. *Psychological Test and Assessment Modeling, 56,* 127-136.

Kubinger, K.D., Heuberger, N., & Poinstingl, H. (2010). On the self-evaluation of a journal's impact factor. *Psychological Test and Assessment Modeling,, 52,* 142-147.

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). *Statistik in der Psychologie – vom Einführungskurs bis zur Dissertation* [Statistics in Psychology – from introductory course through to the doctoral thesis]. Göttingen: Hogrefe.

Leana-Taşcılar, M.Z. (2015). Age differences in the Actiotope Model of Giftedness in a Turkish sample. *Psychological Test and Assessment Modeling, 57,* 111-125.

Niederkofler, B., Herrmann, C., Seiler, S., & Gerlach, E. (2015). What influences motivation in Physical Education? A multilevel approach for identifying climate determinants of achievement motivation. *Psychological Test and Assessment Modeling, 57,* 70-93.

Paz-Baruch, N. (2015). Validation study of the Questionnaire of Educational and Learning Capital (QELC) in Israel. *Psychological Test and Assessment Modeling, 57,* 222-235.

Proyer, R.T., Wagner-Menghin, M.M., & Grafinger, G. (2014). Screening reading comprehension in adults: Development and initial evaluation of a reading comprehension measure. *Psychological Test and Assessment Modeling, 56,* 368-381.

Ranger, J., & Kuhn, J.T. (2014). Testing fit of latent trait models for responses and response times in tests. *Psychological Test and Assessment Modeling, 56,* 382-404.

Rasch, D., Herrendörfer, G., Bock, J., Victor, N., & Guiard, V. (2008). *Verfahrensbibliothek Versuchsplanung und -auswertung. Elektronisches Buch.* [Collection of Procedures in Design and Analysis of Experiments. Electronic Book]. München: Oldenbourg.

Rasch, D., Kubinger, K.D., & Yanagida, T. (2011). *Statistics in Psychology – Using R and SPSS.* Chichester: Wiley.

Rasch, D., Pilz, J., Verdooren, R. L., & Gebhardt, A. (2011). *Optimal experimental design with R*. New York: Chapman & Hall/CRC.

Reutlinger, M., Ballmann, A., Vialle, W., Zhang, Z., & Ziegler, A. (2015). Parental goal orientations for their kindergarten children: Introducing the Nuremberg Parental Goal Orientation Scales (NuPaGOS). *Psychological Test and Assessment Modeling, 57,* 163-178.

Rose, N., von Davier, M., & Nagengast, B. (2015). Commonalities and differences in IRT-based methods for nonignorable item nonresponses. *Psychological Test and Assessment Modeling, 57,* 472-498.

Salzberger, T. (2015). The validity of polytomous items in the Rasch model – The role of statistical evidence of the threshold order. *Psychological Test and Assessment Modeling, 57,* 377-395.

Schneider, B., Rasch, D, Kubinger, K.D., & Yanagida, T. (2015). A sequential triangular test of a correlation coefficient's null-hypothesis: $0 < \rho \leq \rho_0$. *Statistical Papers, 56,* 689-699.

Schweizer, K. (2014). On the ways of investigating the discriminant validity of a scale in giving special emphasis to estimation problems when investigating multitrait-multimethod matrices. *Psychological Test and Assessment Modeling, 56,* 45-59.

Schweizer, K. & Reiß, S. (2014). The structural validity of the FPI Neuroticism scale revisited in the framework of the generalized linear model. *Psychological Test and Assessment Modeling, 56,* 332-347.

Steinbach, J., & Stöger, H. (2015). Measurement of optimal learning environments: Validation of the parents' attitudes towards self-regulated learning scale. *Psychological Test and Assessment Modeling, 57,* 179-200.

Steuer, G., & Dresel, M. (2015). A constructive error climate as an element of effective learning environments. *Psychological Test and Assessment Modeling, 57,* 262-275.

Themessl-Huber, M. (2014). Evaluation of the χ^2-statistic and different fit-indices under misspecified number of factors in confirmatory factor analysis. *Psychological Test and Assessment Modeling, 56,* 219-236.

Tirp, J., Steingröver, C., Wattie, N., Baker, J., & Schorer, J. (2015). Virtual realities as optimal learning environments in sport – A transfer study of virtual and real dart throwing. *Psychological Test and Assessment Modeling, 57,* 57-69.

Vidotto, D., Vermunt, J.K., & Kaptein, M.C. (2015). Multiple imputation of missing categorical data using latent class models: state of the art. *Psychological Test and Assessment Modeling, 57*, 542-576.

Vink, G., Lazendic, G., & van Buuren, S. (2015). partitioned predictive mean matching as a multilevel imputation technique. *Psychological Test and Assessment Modeling, 57,* 577-594.

Vladut, A., Vialle, W., & Ziegler, A. (2015). Learning resources within the Actiotope: A validation study of the QELC (Questionnaire of Educational and Learning Capital). *Psychological Test and Assessment Modeling, 57,* 40-56.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52,* 8-28.

Warne, R.T., & Larsen, R. (2014). Evaluating a proposed modification of the Guttman rule for determining the number of factors in an exploratory factor analysis. *Psychological Test and Assessment Modeling, 56,* 104-123.

Westhoff, K. (2014). The Decision-Oriented Interview (DOI) as an in-depth selection interview. *Psychological Test and Assessment Modeling, 56,* 137-153.

Westhoff, K., & Hagemeister, C. (2014). Competence-oriented oral examinations: objective and valid. *Psychological Test and Assessment Modeling, 56,* 319-331.

Wind, S.A. (2015). Evaluating the quality of analytic ratings with Mokken scaling. *Psychological Test and Assessment Modeling, 57,* 423-444.

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.

Yanagida, T., Strohmeier, D., Toda, Y., & Spiel, C. (2014). The Self Group Distinction Scale: A new approach to measure individualism and collectivism in adolescents. *Psychological Test and Assessment Modeling, 56,* 304-313.

Zhou, X., & Reckase, M.D. (2014). Optimal item pool design for computerized adaptive tests with polytomous items using GPCM. *Psychological Test and Assessment Modeling, 56,* 255-274.