Psychological Test and Assessment Modeling, Volume 63, 2021 (1), 3–11

# Editorial: Necessity for in-depth research work on Psychological Test and Assessment Modeling

*Klaus D. Kubinger (Editor in chief)*

## Introduction

This journal *Psychological Test and Assessment Modeling*, focusses on three areas: psychology-specific statistical methods & problems, general psychometrics, and psychological assessment in theory & practice. Although, there have been a lot of papers published in this journal, which helped to advance certain topics in these areas, further research work is necessary.

Therefore, in the following we will be looking into such topics. These may be subjectively selected, and should hence, at any rate, be broadened. Researchers are encouraged to do so – particularly in this journal. The topics dealt with in the following, refer, in case, to previous papers of *Psychological Test and Assessment Modeling*. In doing so we emphasize the scientific significance of this journal. Regrettably, although some of these papers offer at least first steps towards solving a certain problem, they fail to be noticed in the scientific community, as a consequence they hardly have an impact. For this, it sounds worthwhile to give a respective reminder, here.

## Necessity for in-depth research work on psychology-specific statistical methods and problems

A necessary reminder concerns the practice of rating significant results with a different amount of asterisks, when testing a certain null-hypothesis. Although obvious to graduated statisticians, for instance D. Rasch, Kubinger, Schmidtke, and Häusler (2004, p. 227) outlined in this journal "that the practise of using one, two, or three asterisks (according to a type-I-risk $\alpha$ either 0.05, 0.01, or 0.001) in significance testing … is in no way in accordance with the *Neyman-Pearson* theory of statistical hypothesis testing. Claiming a-posteriori that even a low type-I-risk $\alpha$ leads to significance merely discloses a researcher's self-deception." In fact, "the 'practice of asterisks' always implies the highest $\alpha$ from all $\alpha$-levels that one would ever accept: If a researcher decides, according to the result, what level of $\alpha$ he/she applies (in order to get a significant result that might

be even just at $\alpha = .05$), then the general $\alpha$-level is one that would suffice even in the worst case." (p. 232). That is, if some low type-I-risk is indeed relevant then this one must be definitely declared in advance. However, it is most likely that many researchers only try to convey their results' conclusiveness to the reader. But even for research work in psychology it has become almost the standard to quote the estimated effect of a significant result instead. Furthermore, if any effect is actually relevant and occurs to be significant with a type-I-risk which is lower than fundamentally sufficient, then the sample size of the empirical study proves just too large.

This conclusion is due to the approach of "planning a study". That is, the needed sample size might be calculated in advance, by determining a certain type-I-risk ($\alpha$) and a certain type-II-risk ($\beta$) as well as a relevant effect size – the latter for instance concerning the effective difference of the variable's means in different populations, or concerning the effective strength of relationship of two variables. In doing so, a significant result indicates a relevant one. (For an introduction into this approach on a psychology-researcher level see for instance D. Rasch, Kubinger, & Yanagida, 2011; most favorable is to use the R-routine OPDOE [*OPtimal Design Of Experiments*], D. Rasch, Pilz, Verdooren, & Gebhardt, 2011). Admittedly, planning a study is currently only for parametric tests and almost only for univariate analyses at a researcher's disposal. Hence, further research is needed.

The same is true with so-called sequential testing. Thereby data are sampled one after the other until either the null- or the alternative hypothesis is to accept and the other to reject. This approach is preferable over planning a study alone, because it generally saves quite a lot of needed sample size.

All these considered approaches are particularly relevant for *Pearson*'s correlation coefficient. The problems relating to planning a study (cf. for instance D. Rasch, Kubinger, & Yanagida, 2011) and relating to sequential testing have indeed been solved (see below). But common practice of research work in psychology implies a substantial problem. There, testing almost always refers only to the null-hypothesis $H_0$: $\rho = 0$. If this hypothesis is to reject, the relationship of the two variables in question is then labelled "significant", indicating some quality criterion like a (relevant) effect size. However, obviously a "significant" correlation coefficient is almost meaningless because even a correlation coefficient of .01 can reach significance, given a large enough sample size. Hence, as for instance Kubinger, Rasch, and Šimečkova (2007, in this journal) pointed out, rather the null-hypothesis $H_0$: $0 < \rho \leq \rho_0$ (e.g. $\rho_0 = .70$) should be tested than $H_0$: $\rho = 0$. Schneider, Rasch, Kubinger, and Yanagida (2015; see also D. Rasch, Yanagida, Kubinger, and Schneider, 2018) established a sequential (triangular) test of a correlation coefficient's null-hypothesis $H_0$: $0 < \rho \leq \rho_0$ – for a computer program see the R-routine *seqtest* (Yanagida, 2016). That is: Papers applying this approach may serve as a stimulation for other researchers, towards actual conclusive results being based on a minimal sample size.

Finally, attention shall be called to basic statistical problems. All above researchers

especially pay too much attention to the possible violation of the assumption of the variables' normal distribution which is needed to derive the exact distribution of the statistic in question. But D. Rasch and Guiard (2004) – with 254 citations according to Google Scholar (2021-03-23), perhaps making it the most cited paper of this journal – established a turning point: "in most practical cases the parametric approach for inferences about means is so robust that it can be recommended in nearly all applications" (p. 175). In particular for *Pearson*'s correlation coefficient Yanagida, Rasch, Kubinger, and Schneider (2017) investigated the respective robustness. And D. Rasch, Kubinger, and Moder (2011) additionally proved that for the two-sample $t$-test pre-testing its assumption of homogeneity of variances "does not pay" off. With respect to homogeneity of variances Moder (2010) offers an alternative of the conventionally used $F$-test in one-way analysis of variance (see also D. Rasch, Kubinger, & Yanagida, 2011); however, for more-way analyses of variance the problem of heterogeneity of variances seems still not properly solved; furthermore this is true for multivariate approaches.

## Necessity for in-depth research work on psychometrics

As concerns *Item Response Theory* (IRT), models extending the number of item parameters beyond the Rasch model are evidently of broad interest. However, it seems worthwhile to point out, that such "generalizations", that is modelling more than a single item parameter (i.e. an item difficulty parameter), lose sight of the Rasch model's exceptional property from the perspective of the theory of science: The Rasch model allows specific objective comparisons of examinees or items (see in particular Scheiblechner, 2009, in this journal). This implies measurements that put objects, e.g. examinees, in empirically adequate relationships to one another, irrespectively of other objects which have ever been or will further be considered. This characteristic of the Rasch model ensures that there are means of testing the model: given empirical data, the hypothesis of whether the model holds for a certain item pool can be tested. Be aware that contrary to this, pertinent goodness-of-fit indices, which are applied to many other models, only indicate the extent to which the given data can be explained by the model. That is, in the case of the Rasch model it concerns the (absolute) validness of the model and not only the goodness of fitting the given data, possibly in comparison to other competing models. To conclude: From the perspective of the theory of science, in-depth research work on the Rasch model seems still desirable.

Many researchers refer in their Rasch model applications only to the concept of goodness-of-fit, while they ignore the availability of model tests. This is supported by parameter estimation procedures other than Rasch's (G. Rasch, 1960/1980) based on conditional maximum likelihood estimation. However, there are several well-known approaches of testing the Rasch model referring to the concept of "specific objectivity" (for an overview, see Glas & Verhelst, 1995; Kubinger, 1989; – most of them are implemented in the R-packages *eRm*: Extended Rasch Modeling, Mair, Hatzinger, & Maier, 2015; and *tcl*: Testing in conditional likelihood context, Draxler & Kurz, 2019). And although even

new approaches are arising – all above that of a tree-based method by Strobl, Kopf, and Zeileis (2015) –, it suffers from lack of essential developmental work. There hardly exists a study (apart from Futschek, 2014, in this journal), or even better, a mathematical proof for the respective test-statistic's actual distribution and hence for the type-II-risk which has to be taken into account when applying such statistical tests. The latter goes along with too little consideration on how the respective effect size should be determined. As a consequence, planning a study in order to calculate the necessary sample size in advance (see above) is rather rare. There are first approaches (see Draxler & Kubinger, 2018). While Draxler (2010) and Draxler and Alexandrowicz (2015) base their mathematical derivation on some Wald-statistic, Kubinger, Rasch, and Yanagida (2009, in this journal; 2011) suggest the use of a three-way hierarchical mixed-model analysis of variance – in doing so a new model test of the Rasch model has been established; this approach for calculating the sample size even works with missing values due to the use of test-booklets (Yanagida, Kubinger, & Rasch, 2015). Nevertheless, more knowledge is needed of the actual type-I- and type-II-risk, given a relevant effect size and a certain sample size when testing the Rasch model. And concerning the application of the concept of sequential testing (see above), this seems far from a solution.

Furthermore, any progress of (Rasch model) parameter estimation is relevant. For instance Heine and Tarnai (2015, in this journal) reactivated the approach of pairwise conditional item comparison in order to meet the case of missing data by chance – additionally, Hohensinn and Kubinger (2011, in this journal) proved (again) that scoring omitted items as incorrect instead of missing by system leads to seriously biased item parameters. And Zwitser and Maris (2015) solved, to some extent, the problem that pertinent Rasch model item calibration leads to biased item parameter estimations if items are administered using "multi-stage" testing. This is of great importance for branched adaptive testing; it has to be noted, however, that the problem of biased item parameter estimation is not solved for tailored testing.

Calibrating an item pool according to the Rasch model usually means that non-fitting items are (step-wise) eliminated. Because this is leading, in the best case, only to an *a-posteriori* model validness, sometimes a "kind of cross-validation" (D. Rasch, Kubinger, & Yanagida, 2011) takes place: Given the remaining item pool proves actually to fit the model, then data sampled later are used in order to confirm the model's validness for that pool. This approach needs (simulation) studies which give evidence of the type-II-error in the case that such "kind of cross-validation" is neglected. But for the calibration of an item pool according to the Rasch model there might be an alternative or additional approach. One can think of the detection of model-counter testees: for example people who do not fit the model because of being fatigued or careless, or cheating or guessing on items. Artner (2016, in this journal) proved in his simulation study that removing testees who appear suspicious, according to some person-fit index, does seriously increment the specifity of the model test under question. That is, in practice, item pools might be assessed as contradicting the Rasch model while only (a lot of) testees were used for calibration who do not behave model-conform. Again further studies are needed in order

to establish some generalized recommendations on how to optimally process if item calibration shall take model-counter testees into account.

Indeed, in the context of the so-called LLTM (Linear Logistic Test Model; see for instance Fischer, 2005) no psychometric problem arises. This model, a specific modification of the Rasch model, decomposes the item parameters of the Rasch model by a linear combination of some hypothesized elementary operation parameters. Kubinger (2008, in this journal; see also 2009) illustrates the broad field of applications, from constructing tests using item generating rules to measuring item administration effects. Regrettably, although there are some published applications, particularly in this journal (Effatpanah & Baghaei, 2021; Sonnleitner, 2008), this model is not taken advantage of enough in psychological test construction. Yet, Kubinger, Hohensinn, Holocher-Ertl, and Heuberger (2011, in this journal) have tried for some "inverse" LLTM, that is the decomposition of the person parameters instead of the item parameters of the Rasch model.

With regard to factor analysis as a psychometric means there are two concerns. First, there still are psychological tests in everyday use which are based on factor analysis applied for dichotomous data – despite advised caution for instance by Kubinger (2003, in this journal): Factor analysis in the case of dichotomous variables often leads to artificial factors, that is the resulting factors correspond primarily to different levels of item difficulty. But simply analyzing tetrachoric correlation coefficients instead of conventionally used *Pearson* correlation coefficients overcomes this problem. Most likely only additional studies, all above simulation studies, would serve to respective awareness of test developers. Second, especially with respect to confirmatory factor analysis the pertinent interpretation rules of several goodness-of-fit indices seem challenging. So far, Themessl-Huber (2014, in this journal) found in his simulation study that some of them are proper, others are not.

## Necessity for in-depth research work on psychological assessment

According to the original standards of DIN 33430 (Deutsches Institut für Normung e. V. [DIN], 2002) which is the basis for the world-wide known ISO 10667 (*Assessment service delivery — Procedures and methods to assess people in work and organizational settings;* ISO, 2011a, 2011b) the up-to-datedness of the standardization of a psychological test must be proven every eight years – though the number of years, eight, is completely arbitrary, this demand is reasonable alone regarding the so-called Flynn effect (cf. Flynn, 1996; notice that there are nowadays results which indicate the respective trend just in the opposite direction; Dutton, van der Linden, & Lynn, 2016). Consequently, test-authors and test-publishers are obliged to recurrently check their tests' standardization. However, this is very expensive, as standardization of psychological tests customarily requires about 2000 representatively sampled testees. In particular with tests provided for individual- but not for group-testing, this would mean a vast undertaking. For this, methods are needed to minimize this effort. Sequential testing (see above) would certainly be the method of choice, although it is rather a question of the optimal sampling method

regarding to the representativity of certain sub-populations than the statistical approach. Yet, such a concept does not exist.

Furthermore, standardization of psychological assessment instruments very often lacks from random sampling. Instead, people who are easily at a test author's disposal and are willing to give consent to be tested are used. This practice of taking on these so-called volunteers leads to the problem of not being representative: Volunteers may differ essentially from the targeted population regarding achievement motivation and ability as well as regarding their personality. Hence, if there were diverse non-responder analyses (cf. e.g. Kubinger, 2019) then authors, publishers, and users would have at least a framework for judging standardization's representativity.

Not only adequacy of the standardization of a psychological test is in question when the test is adapted for another culture/language; but also its scaling of calibration is, indeed: Above all, the model-fit according to IRT must be confirmed before a test is applicable to a new population other than the source population previously used for calibration. Meaning, every such test adaption needs an equivalency check. Again, thorough experience of those conditions, which make adaptions more or less likely to stand this equivalency check (see for instance the adaption of the original German *Adaptive Intelligence Diagnosticum*, AID, as an English version; Kubinger, 2017), is necessary.

Commonly, group-testing within psychological assessment uses time limits for working on items which entails a serious problem: Speed-and-power combined achievements cannot claim to measure uni-dimensionally without empirical evidence. And as indicated above, Hohensinn and Kubinger (2011, in his journal) proved that, scoring items as incorrect if a testee does not get to them, underestimates his/her ability parameter gravely. But there are two IRT-based options in order to avoid contamination of power with speed. First, only the items the testee actually worked on are scored. Second, simply as much time for a (sub-) test is provided, as the slowest testee of the group needs until he/she has worked on a defined minimum number of items – both options are covered by the adaptive testing approach! And both are realized within the intelligence test-battery AID-G (*Intelligence Diagnosticum for Group administration*; Kubinger & Hagenmüller, 2019). Further proofs of their practical usefulness is desirable.

**Postscript:** We welcome paper dealing especially with research reproducibility according to the standards set by Hothorn and Leisch (2011). Also authors are warmly encouraged to publish new computer routines (particularly done in R), which support *Psychological Test and Assessment Modeling*.

## References

Artner, R. (2016). A simulation study of person-fit in the Rasch model. *Psychological Test and Assessment Modeling*, *58*, 531–563.

Deutsches Institut für Normung e. V. (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen [Requirements on precedures for the assessment of professional aptitude]*. Berlin: Beuth.

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.

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.

Draxler, C., & Kurz, A. (2019). *tcl: Testing in Conditional Likelihood Context*. R package version 0.1.0. Retrieved from https://CRAN.R-project.org/package=tcl

Dutton, E., van der Linden, D., & Lynn, R. (2016). The negative Flynn Effect: A systematic literature review. *Intelligence*, *59*, 163–169.

Effatpanah, F., & Baghaei, P. (2021). Cognitive components of writing in a second language: An analysis with the Linear Logistic Test Model. *Psychological Test and Assessment Modeling*, *63*, 13–44.

Fischer, G. H. (2005). Linear logistic test models. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 505–514). New York: Elsevier.

Flynn, J. R. (1996). What environmental factors affect intelligence: The relevance of IQ gains over time. In D. K. Detterman (Ed.), *Current topics in human intelligence: The environment* (pp. 17–29). Westport: Ablex Publishing Corporation.

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.

Glas, 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: Springer.

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

Hohensinn, C., & Kubinger, K. D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling*, *53*, 380–393.

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

ISO. (2011a). *ISO 10667-1:2011: Assessment service delivery — Procedures and methods to assess people in work and organizational settings — Part 1: Requirements for the client*. Genf: ISO.

ISO. (2011b). *ISO 10667-2:2011: Assessment service delivery — Procedures and methods to assess people in work and organizational settings — Part 2: Requirements for service providers*. Genf: ISO.

Kubinger, K. D. (1989). Aktueller Stand und kritische Würdigung der Probabilistischen Testtheorie [Critical evaluation of latent trait theory]. In K. D. Kubinger (Ed.), *Moderne Testtheorie – Ein Abriß samt neuesten Beiträgen [Modern psychometrics – A brief survey with recent contributions]* (pp. 19–83). München: PVU.

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, *45*, 106–110.

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. (2009). Applications of the Linear Logistic Test Model in psychometric research. *Educational and Psychological Measurement*, *69*, 232–244.

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

Kubinger, K. D. (2019). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens [Psychological Assessment: Theory and Practice of Psychological Consulting]* (3rd ed.). Göttingen: Hogrefe.

Kubinger, K. D., & Hagenmüller, B. (2019). *Gruppentest zur Erfassung der Intelligenz auf Basis des AID (AID-G) [Intelligence diagnosticum for group administration]*. Göttingen: Hogrefe.

Kubinger, K. D., Hohensinn, C., Holocher-Ertl, S., & Heuberger, N. (2011). Applying the LLTM for the determination of children's cognitive age-acceleration function. *Psychological Test and Assessment Modeling*, *53*, 183–191.

Kubinger, K. D., Rasch, D., & Šimečkova, M. (2007). Testing a correlation coefficient's significance: Using $H_0$: $0 < \rho \leq \lambda$ is preferable to $H_0$: $\rho = 0$. *Psychology Science*, *49*, 74–87.

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.

Mair, P., Hatzinger, R., & Maier, M. J. (2015). *eRm: Extended Rasch Modeling*. 0.15-5. Retrieved from https://cran.r-project.org/package=eRm

Moder, K. (2010). Alternatives to $F$-Test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling*, *52*, 343–353.

Rasch, D., & Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, *46*, 175–208.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample $t$-test: Pre-testing its assumptions does not pay off. *Statistical Papers*, *52*, 219–231.

Rasch, D., Kubinger, K. D., Schmidtke, J., & Häusler, J. (2004). The misuse of asterisks in hypothesis testing. *Psychology Science*, *46*, 227–242.

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.

Rasch, D., Yanagida, T., Kubinger, K. D., & Schneider, B. (2018). Determination of the optimal size of subsamples for testing a correlation coefficient by a sequential triangular test. In J. Pilz, D. Rasch, V. B. Melas, & K. Moder (Eds.), *Statistics and simulation* (pp. 315–328). Heidelberg: Springer.

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

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

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.

Sonnleitner, P. (2008). Using the LLTM to evaluate an item generating system for reading comprehension. *Psychology Science Quarterly*, *50*, 345–362.

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*, 289–316.

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

Yanagida, T. (2016). *seqtest: Sequential Triangular Test*. R package version 0.1-0. Retrieved from https://CRAN.R-project.org/package=seqtest

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., Rasch, D., Kubinger, K. D., & Schneider, B. (2017). Robustness of the test of a product moment correlation coefficient under nonnormality. *Journal of Statistical Theory and Practice*, *11*, 493–502. https://doi.org/10.1080/15598608.2017.1329102

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