Psychological Test and Assessment Modeling, Volume 56, 2014 (2), 127-136

Editorial: Toward essential contributions for Psychological Test and Assessment Modeling

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

Preamble: This journal, focusing on "Psychological Test and Assessment Modeling" since 2010, has received great attention from researchers interested in psychology-specific statistical methods and problems, general psychometrics, and psychological assessment in theory and practice. That is, a journal with deep methodical approaches enlarges the scene. Of course, journals of methods in psychology (and educational science) suffer from the lack of high impact factors, on average, and from the lack of willing competent reviewers. While the matter of impact factors is actually encouraging, though fluctuating – according to Kubinger, Heuberger, and Poinstingl (2010) the self-evaluated impact factor is 2010: 0.565, 2011: 0.525, 2012: 0.783, 2013: 0.420 –, reviewers are rare. Nevertheless, we aim for a very quick processing of submitted papers, giving precise reviews of how to improve the quality of a paper – if worthwhile. In the following editorial we once again outline the scope of the journal, but primarily give hints on how to manage research work in order to contribute to the concerning area by a very high methodical standard.

Statistical standards

Within psychology there are several misuses of statistical analyses, at least some improper traditions. Rasch, Kubinger, and Yanagida (2011), for instance – if the reader prefers German, see Kubinger, Rasch, and Yanagida (2011) – elucidate many of them.

First of all, it is to be emphasized that the type-I-error (the significance level) must be established in advance, but not when some p-value has indeed been calculated: For instance also in the predecessor of this journal, Rasch, Kubinger, Schmidtke, and Häusler (2004, p. 232) outlined, that "the 'practice of asterisks' always implies the highest α from all α-levels that one would ever accept: If a researcher decides, according to the result, what level of α he/she applies (in order to get a significant result that might be even just at α=.05), then the

[1] *Correspondence concerning this article should be addressed to:* Prof. 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

general α-level is one that would suffice even in the worst case. Claiming a-posteriori that a lower α applies, would merely disclose the researcher's self-deception."

In the second instant, obviously a proper approach would be to argue in detail, why a certain type-I-risk (α) is established, all above with respect to the practical consequences of an eventual type-I-error in relation to an eventual type-II-error. It would be even better to plan the study in advance regarding the necessary sample size: Given a certain type-I- and type-II-risk as well as a relevant effect size, pertinent computer programs, especially the R-routine OPDOE (*OPtimal Design Of Experiments*); Rasch, Pilz, Verdooren, & Gebhardt, 2011) calculate the sample size so that only relevant effects will result in significance, but such relevant effects will not be detected only with a probability of the type-II-risk settled. Admittedly, planning a study in such a way is currently only at a researcher's disposal for parametric tests such as the two-sample *t*-test and the analysis of variance, and moreover almost only for univariate analyses – "Planning the study as concerns non-parametric methods generally results in problems because the alternative hypothesis is hard to quantify." (Rasch, Kubinger, & Yanagida, 2011, p. 216). However, planning according to some parametric test at least, although a non-parametric homologous test is aimed-for, rather meets the statistical state-of-the-art than fixing the sample size arbitrarily with the consequence, that often significant, but irrelevant effect sizes result or the other way round, that no significance occurs although the estimated effect size is indeed of practical relevance. Concerning multivariate analyses in general, Rasch, Kubinger, and Yanagida (2011, p. 418) emphasize: "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)." A more satisfying situation is given particularly for the discriminant analysis: 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 the given precision requirements type-I-, type-II-risk, and effect size (see Rasch, Herrendörfer, Bock, Victor, & Guiard, 2008).

Thirdly, it would be even better to apply the approach of sequential testing, which is especially attractive if the data are sampled one after the other. Then the data observed so far can be analysed before any next research unit will be sampled. Given the same precision requirements as above, there are three instead of just two decision options, thus either to accept or to reject H_0, or to continue the study with data sampling. The advantage of sequential testing compared to the "classical" approach is that, on average, much less research units are needed. The R-routine OPDOE serves for such analyses as well. Admittedly again, sequential testing is only at a researcher's disposal for parametric tests and moreover not for multivariate analyses.

A fourth issue meets the rumor that normal distribution and homogeneity of variances is an indispensable presupposition for several parametric tests. In an early paper, also in the predecessor of this journal, Rasch and Guiard (2004) demonstrated that the two-sample *t*-test and analysis of variance is at least 20% robust (meaning the actual type-I-risk is for instance in between 4% and 6% when the nominal risk is 5%), even if the distribution of the data differs extremely from the normal distribution (see also von Eye, 2004; and Guiard & Rasch, 2004; both in the predecessor of this journal, as well). In the meantime Rasch, Kubinger, and Moder (2011) proved, as concerns the two-sample *t*-test, that pre-testing the theory-based assumptions of normal distributions and homogeneity of the variances leads to unknown final type-I- and type-II-risks if the respective statistical tests are performed using the same set of observations – which is usually the case. As a consequence of their simulation studies, they recommend "to apply no pre-tests for the *t* test and no *t* test at all, but instead to use the Welch-test as a standard test: its power comes close to that of the *t* test when the variances are homogeneous, and for unequal variances and skewness values $|\gamma_1| < 3$ it keeps the so called 20% robustness whereas the *t* test as well as Wilcoxon's *U* test cannot be recommended for most cases." (p. 219). And Moder (2010) presented the results of a simulation study in this journal, which discloses that it is best to apply Hotteling's T^2 instead of the analysis of variance if homogeneity is doubtful – any pre-test is not worthwhile, because of its low power. In the case of multivariate analyses, however, homogeneity of co-variances are essential and Box's M-test is a *sine qua non*; if this test results in significance, it is just recommended to delete the sample (group) or the variable which is responsible for this. Concerning the assumption of multivariate normal distribution, the matter is that all the tests at hand only have insufficient power (cf. von Eye & Bogat, 2004, also in the predecessor of this journal) – this includes the well-known Kolmogorov-Smirnov test of uni-dimensional normal distribution as well. Of course, a multiple test for uni-dimensional normal distribution does not suffice at all.

Fifth, a great misunderstanding seems to be established in psychology, as significant correlation coefficients are very often interpreted as a "measure of all things", ignoring that only the determination coefficient, which is the squared (Pearson) correlation coefficient (and multiplied by 100), gives conclusive information, that is, the percentage of mutually explained variance of both variables under consideration. Of course, this is trivial; but just testing the null-hypothesis, whether a correlation coefficient differs in the population from 0, feeds overemphasizing the magnitude of correlation. Obviously, even a correlation coefficient of .01 can reach significance, given the sample size is large enough, but there is no meaning of a correlation of that magnitude. Again, this is a matter of planning the study, in other words, calculating 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. Even more, it is a question of null-hypothesis: Instead of testing H_0: $\rho = 0$ it is preferable to test H_0: $0 < \rho \leq \rho_0$, e.g. $\rho_0 = .7$ (in order to use even SPSS for such a test Kubinger, Rasch, & Šimečkova, 2007, provided – also in the predecessor of this journal – a respective syntax); when the latter null-hypothesis is to be rejected, then we have the information that within the population the correlation is at least so strong that $\rho_0^2 \cdot 100\%$ of the variance is mutually explained (in the example that is 49% ≈ 50%). Certainly, if such a test is planned in advance, which again is offered particularly by the R-routine OPDOE, we would additionally be sure that failing to discover such an effect (ρ_0^2) happens to occur only with the settled type-II-risk. Even more elaborated would be sequential testing of the null-hypothesis H_0: $0 < \rho \leq \rho_0$: Recently Schneider, Rasch, Kubinger, and Yanagida (in print) published a test, which is on its way to be integrated in the R-routine OPDOE.

Finally, the sixth issue is applying conventional (i.e. SPSS-default) factor analysis on dichotomous data. Of course, there is a lot of evidence, going back to Guttman (1955) at least, that doing so will most likely result in as many factors as there are variables with different marginal distributions. Yet again Kubinger (2003; in the predecessor of this journal, too) pointed out that using the tetrachoric correlation coefficient instead of Pearson's would be the simplest solution, though his respective SPSS-Syntax was then offered just via his homepage. In the meanwhile applying R will close this gap.

Psychometric standards

Due to the fact that large-scale assessment studies – mostly intended to measure as many subject areas and, consequently, as many items as possible – entail the use of test-booklets with partially different items, but the testees must be compared with their test performance nevertheless, it can be said: models of Item Response Theory (IRT), all above the Rasch model and its generalizations became of extraordinary importance (and late acknowledgement); cf. Adams, Wilson, and Wu (1997). Such models exclusively offer the possibility of linking items of different test-booklets in order to calibrate item parameters and consequently comparing testees' ability parameters appropriately. Since then IRT has boomed, that is, the application of its models and fundamental research work on IRT take up a great deal of published papers.

This means in the first instance that psychological (and educational) test calibrations can hardly disregard the findings of IRT. Most notably, this concerns a psychometric law (cf. Fischer, 1995): if the number of solved items is to be considered a sufficient statistic, meaning it serves as the test score, then all items of the test must conform with the Rasch model. Depending on some other concrete scoring rules, there are different models which, if valid, are sufficient for fair and adequate measuring (a summary is given by Kubinger, 1989).

In the second instance, applying any IRT model always needs analyses of how appropriate the model is for the given data. There are model tests on the one hand, which test some model-intrinsic features, as for instance "specific objectivity" of the Rasch model – apart from Rasch's (1960/1980) original work and Fischer (1995) see Scheiblechner (2009; in the predecessor of this journal), who gives a fundamental reflection of what specific objective comparisons mean. If a data set and an item pool, respectively, stands the model test – at best according to Popper's concept of "degree of corroboration/confirmation" (e.g. Popper, 2001) in several independent samples – then model validness is established (cf. for some standards of Rasch model analysis Kubinger, 2005). There are goodness-of-fit indices (sometimes even goodness-of-fit tests) on the other hand, which however principally only compare the data that would be hypothetically expected by a certain model with the data we actually observed; that is, the fit of data and model is evaluated, but the model's validness *per se* is not dealt with. Concerning the software for so-called conditional maximum parameter estimation due to specific objectivity there is now a comfortable R-routine at a researcher's disposal, eRm (*e*xtended

*R*asch *m*odeling; Mair, Hatzinger, & Meier, 2011), which was introduced in the predecessor of this journal (Mair & Hatzinger, 2007).

A third issue is differential item functioning (DIF), meaning that a single item or very few items of an item pool differ with respect to their difficulties' in relation to all the other items between at least two subpopulations. Of course, DIF-analyses can be completed well by models based on specific objectivity. Teresi et al. (2009) gives an excellent overview, again in the predecessor of this journal. Therefore, applying tests for psychological assessment should always be preceded by pertinent DIF-analyses. By the way, the approach of planning a study according to some precision requirements as described above, even works for the Rasch model nowadays: In the predecessor of this journal Kubinger, Rasch, and Yanagida (2009) worked out a proper solution – also see Draxler (2010) for a further approach.

A fourth aspect meets economic testing, which particularly concerns (IRT-based) adaptive testing. As is well known, deliberately choosing items according to a testee's previous performance leads to a smaller number of items needed to administer in order to achieve the same accuracy of measurement as conventional testing (testing with a fixed design of item presentation). Recently, in this journal guest editors were able to be acquired for a special topic dealing extensively with adaptive testing (see Frey & Kröhne, 2012; Kröhne & Frey, 2013); the papers emphasize among other things, that item position effects should be strictly tested in advance (Hartig & Buchholz, 2012; Yousfi & Böhme, 2012; Walter & Rose, 2013; see also Hohensinn et al., 2008, 2011 – the former again in the predecessor of this journal), that (item) parameter estimation may be biased when based on adaptive item administration (Kubinger, Steinfeld, Reif, & Yanagida, 2012; moreover see Eggen & Verhelst, 2011, as well as Zwitser & Maris, 2013), and that multidimensional adaptive testing actually works (Seitz & Frey, 2013; but for more depth also see Frey & Seitz, 2009). All in all, as time resources become a more and more crucial condition for psychological (and educational) testing, adaptive testing should be applied as a standard procedure, rather just occasionally.

For some kind of construct validating a test by using item generating rules and, furthermore, as a means of fundamental research on, for instance, measuring item administration effects, the LLTM ("linear logistic test model"; Fischer, 1973, 2005; see also Kubinger, 2008, in the predecessor of this journal) serves as a fifth issue. There is even a special issue of the predecessor of this journal illustrating the range of application of the LLTM. Just to highlight two contributions going beyond the facilities spoken of, Embretson and Daniel (2008) analysed the complexity (i.e. structural components) of mathematical problem solving items of a widely used large-scale assessment test; and Draney and Wilson (2008) decomposed and separated, respectively, a severity parameter due to the assessors' ratings. Furthermore, in this journal, Kubinger, Hohensinn, Holocher-Ertl, and Heuberger (2011) introduced how the LLTM applies and how pertinent software (e.g. eRm, see above) might even be used, when item and person parameter are interchanged, that is instead of the item parameter, the person parameter will be decomposed by a linear combination of some hypothesized elementary operation parameters. At least at any rate, the LLTM offers a proper means for research on what exactly a psychological (or educational) test measures.

Psychological assessment proceedings

Proceedings in psychological assessment are either based on modeling the interdependencies of personal traits and context variables on the one side, and individual behavior or inner experiencing on the other side; or they are based on some (psycho-) technological elaborations.

Regarding modeling, there have been two issues published in this journal with the special topic of the assessment of giftedness, taking into account that respective identification goes beyond IQ tests (cf. Vialle, 2013a, b). Among other contributions, the topics include identifying the causes of underachievement (Stoeger, Suggate, and Ziegler, 2013), the differentiation of divergent and convergent thinking (Leikin, 2013), the importance of working memory capacity for giftedness (Howard, Johnson, & Pascual-Leone, 2013), and the contribution of personality styles for achieving high performances (Holocher-Ertl, Schubhart, & Wilflinger, 2013) – additional evidence of the relevance of high ability assessment is shown in two special issues of the predecessor of this journal (cf. Stöger & Ziegler, 2008; Ziegler & Stöger, 2004). That is, such modeling does in the long run not only support practitioners in their professional work of psychological assessment, but also exemplifies how to carry out models for describing and predicting complex psychological phenomena in general.

(Psycho-) Technological elaborations concern every effort of improving psychological instruments' validity and accuracy of measurement, the economy and reasonableness of their administration, and of course their fairness – the latter particularly regarding intercultural and globalized effects. Besides the adaptive testing as outlined above, challenging economy and high accuracy of measurement, respectively, and besides the DIF as also discussed above with reference to intercultural effects, the validity is of great concern due to mainly two problems: Fakeability of personality questionnaires and lucky guessing in multiple-choice tests. Although there is not really a solution for both problems, any effort toward minimization of such effects – accompanied by higher reasonableness for the testees – is necessary and again established in some way in this journal or its predecessor: There was even a special issue of "response distortion in personality measurement" (cf. Deller, Ones, Viswesvaran, & Dilchert, 2006); and among other contributions Foster and Miller (2009) introduced a new format for multiple-choice testing, which was even more elaborated by Kingston, Tiemann, Miller, and Foster (2012).

Of course, psychological assessment proceedings may even concern new psychological instruments, measuring "new" constructs; for instance, once more in the predecessor of this journal, Platt, Proyer, and Ruch (2009) dealt with the assessment of gelotophobia. To conclude, any expansion of instruments is of importance if relevant for the professional psychological assessment in our society.

Postscript: Authors are heartily encouraged to contribute to all sketched topics and issues – and to topics and issues going beyond these.

References

Adams, R.J., Wilson, M., & Wu, M. (1997). Multilevel Item Response Models: An Approach to Errors in Variable Regression. *Journal of Educational and Behavioral Statistics, 22,* 47–76.

Draney, K. & Wilson, M. (2008). A LLTM approach to the examination of teacher's ratings of classroom assessment tasks. *Psychology Science Quarterly, 50,* 417-432.

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

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

Embretson, S.E., & Daniel, R.C. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50,* 328-344.

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

Fischer, G. H. (1995). Derivations of the Rasch Model. In G. H. Fischer & I. W. Molenaar (eds.), *Rasch models* (pp. 15–38). New York: Springer.

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

Frey. A., & Kröhne, U. (2012). Special topic: Current issues in Educational and Psychological Measurement: Design, calibration, and adaptive testing (Part 1). Guest Editorial. *Psychological Test and Assessment Modeling, 54,* 363-365.

Frey, A., & Seitz, N. N. (2009). Multidimensional Adaptive Testing in Educational and Psychological Measurement: Current State and Future Challenges. *Studies in Educational Evaluation*, *35*, 89-94.

Guttman, L.A. (1955). A generalized simplex for factor analysis. *Psychometrika*, *20*, 173-192.

Hartig, J., & Buchholz, J. (2012). A multilevel item response model for item position effects and individual persistence. *Psychological Test and Assessment Modeling, 54,* 418-431.

Hohensinn, C., Kubinger, K.D., Reif, M., Holocher-Ertl, S., Khorramdel, L. & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly, 50*, 391-402.

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,* 497–509.

Holocher-Ertl, S., Schubhart, S., & Wilflinger, G. (2013). Intellectual and non-intellectual determinants of high academic achievement – the contribution of personality traits to the assessment of high performance potential. *Psychological Test and Assessment Modeling, 55,* 231-244.

Howard, S.J., Johnson, J., & Pascual-Leone, J. (2013). *Psychological Test and Assessment Modeling, 55,* 250-273.

Kröhne, U., & Frey, A. (2013). Special topic: Current issues in Educational and Psychological Measurement: Design, calibration, and adaptive testing (Part 2). Guest Editorial. *Psychological Test and Assessment Modeling, 55,* 79-80.

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). Munich: PVU.

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

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 composing tests by item generating rules to measuring item administration effects. *Psychology Science Quarterly, 50,* 311-327.

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

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: $0 < \rho \le \lambda$ 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). *Statistik in der Psychologie – vom Einführungskurs bis zur Dissertation* [Statistics in Psychology – from introductory course through to the doctoral thesis]. Göttingen: Hogrefe.

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.

Leikin, R. (2013). Evaluating mathematical creativity: The interplay between multiplicity and insight. *Psychological Test and Assessment Modeling, 55,* 385-400.

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

Mair, P., Hatzinger, R., & Maier M.J. (2014). eRm: Extended Rasch Modeling. R package version 0.15-4 [Computersoftware]. Retrieved from http://erm.r-forge.r-project.org/

Moder, K. (2010). Alternatives to F-Test in One Way ANOVA in case of heterogeneity of variances. *Psychological Test and Assessment Modeling, 52,* 343-353.

Popper, K.R. (2001, reprint). *The logic of scientific discovery*. London: Routledge.

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

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. & 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.

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. (in print). A Sequential Triangular Test of a Correlation Coefficient's Null-Hypothesis: $0 < \rho \leq \rho_0$. *Statistical Papers.*

Seitz, N.N., & Frey, A. (2013). The sequential probability ratio test for multidimensional adaptive testing with between-item multidimensionality. *Psychological Test and Assessment Modeling, 55,* 105-123.

Stoeger, H. & Ziegler, A. (2008). Editorial: High Ability Assessment. *Psychology Science Quarterly, 50,* 91-96.

Stoeger, H., Suggate, S., & Ziegler, A. (2013). Identifying the causes of underachievement: A plea for the inclusion of fine motor skills. *Psychological Test and Assessment Modeling, 55,* 274-288.

Teresi, J. A., Ocepek-Welikson, K., Kleinman, M., Eimicke, J. P., Jones, R. N., Lai, J. S., et al. (2009). Analysis of differential item functioning in the depression item bank from the Patient Reported Outcome Measurement Information System (PROMIS): An item response theory approach. *Psychology Science Quarterly*, *51*, 148-180.

von Eye, A. & Bogat, A. (2004). Testing the assumption of multivariate normality. *Psychology Science, 46,* 243-258.

Vialle, W. (2013a). Special topic: Current perspectives on the assessment of giftedness – Part I. *Psychological Test and Assessment Modeling, 55,* 247-249.

Vialle, W. (2013b). Special topic: Current perspectives on the assessment of giftedness – Part II. *Psychological Test and Assessment Modeling, 55,* 383-384.

Walter, O.B., & Rose., M. (2013). Effect of item order on calibration and item bank construction for computer adaptive tests. *Psychological Test and Assessment Modeling, 55,* 81-91.

Yousfi, S., & Böhme, H.F. (2012). Principles and procedures of considering item sequence effects in the development of calibrated item pools: Conceptual analysis and empirical illustration. *Psychological Test and Assessment Modeling, 54,* 366-396.

Ziegler, A., & Stöger, H. (2004). Identification based on ENTER within the conceptual frame of the actiotope model of giftedness. *Psychology Science, 46,* 324-341.

Zwitser, R.J., & Maris, G. (2013). Conditional statistical inference with multistage testing designs. *Psychometrika,* DOI: 10.1007/S11336-013-9369-6.