Evaluating Testing, Profile Likelihood Confidence Interval Estimation, and Model Comparisons for Item Covariate Effects in Linear Logistic Test Models

Applied Psychological Measurement

2017, Vol. 41(5) 353–371

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621617692078

journals.sagepub.com/home/apm

Sun-Joo Cho¹, Paul De Boeck², and Woo-Yeol Lee¹

Abstract

The linear logistic test model (LLTM) has been widely applied to investigate the effects of item covariates on item difficulty. The LLTM was extended with random item residuals to account for item differences not explained by the item covariates. This extended LLTM is called the LLTM-R. In this article, statistical inference methods are investigated for these two models. Type I error rates and power are compared via Monte Carlo studies. Based on the simulation results, the use of the likelihood ratio test (LRT) is recommended over the paired-sample *t* test based on sum scores, the Wald *z* test, and information criteria, and the LRT is recommended over the profile likelihood confidence interval because of the simplicity of the LRT. In addition, it is concluded that the LLTM-R is the better general model approach. Inferences based on the LLTM while the LLTM-R is the true model appear to be largely biased in the liberal way, while inferences based on the LLTM-R while the LLTM is the true model are only biased in a very minor and conservative way. Furthermore, in the absence of residual variance, Type I error rate and power were acceptable except for power when the number of items is small (10 items) and also the number of persons is small (200 persons). In the presence of residual variance, however, the number of items needs to be large (80 items) to avoid an inflated Type I error and to reach a power level of .90 for a moderate effect.

Keywords

linear logistic test model, model comparison approach, profile likelihood confidence interval, random item residuals, statistical testing

¹Vanderbilt University, Nashville, TN, USA

²Ohio State University, Columbus, OH, USA

Corresponding Author:

Sun-Joo Cho, Vanderbilt University, Peabody Hobbs 213a, 230 Appleton Place, Nashville, TN 37203, USA.

Email: sj.cho@vanderbilt.edu

Introduction

The linear logistic test model (LLTM; Fischer, 1973) has been widely applied when researchers are interested in the effects of item covariates on item difficulty (e.g., Embretson & Wetzel, 1987; Freund, Hofer, & Holling, 2008; Green & Smith, 1987; Hornke & Habon, 1986; Sheenan & Mislevy, 1990; Spada & McGaw, 1985; Whitely & Schneider, 1981). The LLTM was extended to the LLTM with random item residuals (LLTM-R) to account for unexplained item differences, either within a two-stage empirical Bayes regression model framework (Mislevy, 1988) or within a generalized linear mixed effect models (GLMM) framework (De Boeck, 2008; Janssen, Schepers, & Peres, 2004). To determine whether the effects of item covariates in the LLTM and the LLTM-R are statistically significant, null hypothesis significance testing for the effect has been used in most LLTM applications (e.g., Embretson & Wetzel, 1987; Gorin, 2005; Whitely & Schneider, 1981). Test statistics (e.g., *t*- or *z*-statistics) and corresponding *p* values are used to test whether the population value of the effect differs from a specified value (generally 0). In addition to testing the effect, the two models, the LLTM(-R) with and without residual variance, are compared to detect item covariate effects using model comparison approaches such as the likelihood ratio test (LRT) or information criteria (e.g., Hohensinn & Kubinger, 2011; Kubinger, 2009).

The LLTM and the LLTM-R can be formulated as GLMMs (De Boeck, 2008; De Boeck & Wilson, 2004). Null hypothesis significance testing for fixed effects in GLMMs has been addressed using *t*- and *z*-statistics. In applications of the LLTM and the LLTM-R, the challenges are as follows. First, because the number of items is smaller (often less than 50) than the number of persons in many applications, small sample inference may result in inaccurate results when the *z*-statistic is used as a Wald test. For evident reasons, the *z*-statistic can be used only when the number of items is sufficiently large. Second, the *z*-statistic is based on estimated standard errors. Because the estimated standard errors do not take into account the sampling variability introduced by estimating the unknown variance of the estimator, the estimated standard errors underestimate the true variability of fixed effects (Dempster, Rubin, & Tsutakawa, 1981). Although the *t*-statistic may be more appropriate for smaller number of items, it suffers from the same downward bias of the estimated standard error, and in addition it is not clear what the degrees of freedom are. For example, suppose that there are one item covariate and 20 items. When the LLTM-R is applied, the number of item parameters to be estimated is two, that is, one fixed effect of the item covariate, and one variance of random item residuals, and in addition there is a fixed intercept. Does this mean that there are 17 ($= 20 - 3$) degrees of freedom (*df*) for the *t* test, or does also the number of persons count and are the *df* the number of persons times the number of items minus the number of model parameters? And should the *df* not be reduced to compensate for the underestimated standard error? Several methods to calculate *df* have been developed for the fixed effects in GLMM, but the *df* calculation is an ongoing research area in the statistics literature (e.g., Baayen, Davidson, & Bates, 2008, p. 396; Pinheiro & Bates, 2000, pp. 87-92; Molenberghs & Verbeke, 2004, p. 135). Thus, the *t*-statistic based on the fixed effect estimate divided by its estimated standard error was not considered in this study.

As an alternative to null hypothesis significance testing, the use of effect sizes (i.e., estimates of item covariates in the LLTM and the LLTM-R) and their corresponding confidence interval (CI) can be used (e.g., Hecht et al., 2015). CIs can be more useful than null hypothesis significance testing, especially when researchers address issues that involve the magnitude of the effects (Steiger, 2004). In such a case, it is important to obtain sufficiently narrow CIs, which can be evaluated with the accuracy in parameter estimation (Kelly & Maxwell, 2003). The Wald CI works properly when the distribution of the parameter estimator is symmetric,

and the standard error is a good estimate of the standard deviation of the estimator (Wald & Wolfowitz, 1939). These are the same conditions as for the z test. Because the standard error is obtained based on asymptotic variances obtained from the information matrix in maximum likelihood estimation, the Wald CI may not perform well for small sample sizes for the same reason as a z test. Profile likelihood CIs are recommended for small sample sizes because such CIs do not assume the normality of the estimator (e.g., Cox & Hinkley, 1974). To the authors' knowledge, the degree of improvement in using a profile likelihood CI when the LLTM and the LLTM-R are applied to a small number of items (e.g., 10) has not been shown.

Regarding model comparison approaches, Weirich, Hecht, and Böhme (2014) examined the Type I error rate and the power of the LRT when using the LLTM-R for detecting the (fixed) effects of item position effects in a large-scale assessment such as the National Assessment of Educational Progress (NAEP). They found that the LLTM-R provided adequate power (e.g., $>.90$) when there are a large number of persons (e.g., 4,000) and a large number of items (e.g., 80). However, the performance of other model comparison approaches such as the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) has not been investigated for the LLTM and the LLTM-R.

Table 1A in the Online Appendix gives a descriptive overview of LLTM and LLTM-R studies. Item covariate effects have been used to examine the effects of cognitive variables on item difficulty (e.g., Fischer, 1973; Whitely & Schneider, 1981), the effects of item characteristics on item difficulty (e.g., De Boeck, 2008), to design tests based on cognitive rules (e.g., Hornke & Habon, 1986; Gorin, 2005), and to investigate the design effects in large-scale tests (e.g., Hartig, Frey, Nold, & Klieme, 2012; Hecht, Weirich, Siegle, & Frey, 2015). As one can see, various approaches have been used for inferences and no consensus has developed. An additional complication is that estimates and the standard errors of the LLTM can be largely underestimated if random residuals are ignored (e.g., Hartig et al., 2012).

Therefore, the purpose of this study is to address the inferential qualities (Type I error rate and power) of different approaches: null hypothesis significance testing based on the estimates of covariate effects, profile likelihood CI, and model comparisons approaches (i.e., the LRT and information criteria). Although one may also be interested in profile likelihood CIs for other reasons than significance testing, such as effect size considerations, the former was focused on. Specifically, this study investigated whether the inferences hold nominal Type I error rates ($\alpha = .01$ or $\alpha = .05$ or $\alpha = .10$) and whether the power is sufficiently high ($.90$). In addition, the impact of ignoring random residuals was investigated systematically. This article fills a void in the LLTM-related literature.

In the current study, a prototypical case with one binary covariate and a balanced design was chosen. Theoretically a very extensive study is possible, with different types of item covariates (binary-, ordered-, or nominal-category covariates), with different designs (balanced, slightly unbalanced, or largely unbalanced), with different numbers of covariates, and with and without collinearity between covariates. Here, a simple design was considered for several reasons. First, in the more traditional approach, the analogue of the LLTM is a repeated-measures design with a within-subjects factor, and a balanced design is the more prototypical case. For binary data, one would then commonly use sum scores for the two levels of the repeated measures factor, because binary variables are difficult to handle in a traditional approach. In this case, using a paired-sample t test for the sum scores (the equivalent of a simple within-subjects repeated-measures analysis of variance [ANOVA]) would be the way to go to test the effect of the factor. For a similar reasoning, see Kubinger (2009). This allows us to make a link and a useful comparison between two very different approaches: an LLTM approach and a traditional paired-sample t test. Second, Johnson, Barry, Ferguson, and Müller (2015) have found that the type of covariate is not a factor that affects power, and other influences such as correlation

between covariates seem very similar to those in multiple regression (Green & Smith, 1987). There are no special reasons why well-known effects in multiple linear regression (such as the effect of collinearity) would not play a role in the LLTM(-R). For the present study, the focus is not on the types of covariate configurations but on a comparison of inferential approaches, and to begin with investigating this issue, it is useful to concentrate on a simple to interpret case that would not imply various complications and a huge number of conditions in the simulation design. As will become clear from the following, there are already very many comparisons to make for the simple case. Third, because in practice, often more than one item covariate is used, a real data example with multiple covariates was provided, to make the contribution less abstract and to see the relevance for a broader situation than covered in the simulation study.

In the following, the LLTM and the LLTM-R and their parameter estimation methods are described. Next, null hypothesis significance testing, profile likelihood CI, and model comparison approaches are presented for inferences regarding an item covariate effect. Then follows the simulation study: design, hypotheses, and results. The authors end with a summary and a discussion.

LLTM and Parameter Estimation

The LLTM is a constrained Rasch model, and the difficulty parameters of the LLTM are estimated as linear combinations of a smaller number of item covariates such as cognitive processing demands of tasks and item positions in a test. The LLTM is described as follows:

$$\text{logit}[Pr(y_{pi} = 1 | \theta_p)] = \theta_p - \left(\mu + \sum_{d=1}^D \gamma_d Q_{id} \right), \quad (1)$$

where p is a person index ($p = 1, \dots, P$), i is an item index ($i = 1, \dots, I$), d is a covariate index ($d = 1, \dots, D$), y_{pi} is an item response from a person p to item i , Q_{id} is the item covariate d value for item i , θ_p is the normally distributed underlying person trait, $\theta_p \sim N(0, \sigma_\theta^2)$, μ is the overall mean if the item covariates are centered or effect coded, and γ_d is the effect of the item covariate d .

The LLTM-R is expressed in the following equation:

$$\text{logit}[Pr(y_{pi} = 1 | \theta_p, \varepsilon_i)] = \theta_p - \left(\mu + \sum_{d=1}^D \gamma_d Q_{id} + \varepsilon_i \right), \quad (2)$$

where ε_i is the normally distributed residual over items, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$.

Parameter Estimation

In this study, maximum likelihood estimation implemented in glmer of the lme4 R package (Bates, Maechler, & Bolker, 2011) was chosen to estimate the parameters of the LLTM and the LLTM-R. The LLTM-R has crossed-random effects (i.e., θ_p and ε_i), which requires a high-dimensional integration. To avoid the high-dimensional integration, Laplace approximation (corresponding to one adaptive quadrature point) is implemented for the crossed-random effect models in lme4. It has been shown that the (approximate) maximum likelihood estimation provides accurate estimates for the crossed-random effect model (e.g., LLTM-R), although it leads to a downward bias for variance estimates of the random effects (e.g., $\hat{\sigma}_\varepsilon^2$) when cluster sizes (e.g., the number of items) are small (e.g., 10 items) (Chalmers, 2015; Cho & Rabe-Hesketh, 2011; Joe, 2008).

Because the model comparison methods chosen in this article (LRT and information criteria) use maximized log-likelihood values, it may be instructive to present marginalized likelihood functions for the LLTM and the LLTM-R. The marginal likelihood for the LLTM is as follows:

$$\int_{\theta} \prod_{pi} Pr(y_{pi} = 1 | \theta_p)^{y_{pi}} [1 - Pr(y_{pi} = 1 | \theta_p)]^{1-y_{pi}} \prod_p g_1(\theta_p) d\theta, \quad (3)$$

where $g_1(\theta_p)$ is a normal density with a zero mean and a standard deviation equal to σ_{θ} and $\theta = [\theta_1, \dots, \theta_p, \dots, \theta_P]$. The marginal likelihood for LLTM-R is as follows:

$$\int_{\varepsilon} \int_{\theta} \prod_{pi} Pr(y_{pi} = 1 | \theta_p, \varepsilon_i)^{y_{pi}} [1 - Pr(y_{pi} = 1 | \theta_p, \varepsilon_i)]^{1-y_{pi}} \cdot \prod_p g_1(\theta_p) \cdot \prod_i g_2(\varepsilon_i) d\varepsilon d\theta, \quad (4)$$

where $g_2(\varepsilon_i)$ is the normal density with a zero mean and a standard deviation equal to σ_{ε} and $\varepsilon = [\varepsilon_1, \dots, \varepsilon_i, \dots, \varepsilon_I]$.

Detecting Item Covariate Effects

In this section, the null hypothesis significance testing, profile likelihood CI, and model comparison approaches are described for detecting γ_d in the LLTM and the LLTM-R (Equations 1 and 2, respectively) and the sum-based paired-sample t test is also described.

Null Hypothesis Significance Testing

z test for γ_d . For the effect of an item covariate, γ_d , an approximate z test is obtained from approximating the distribution of $(\hat{\gamma}_d - \gamma_d)/SE_{\hat{\gamma}_d}$ by a standard univariate normal distribution. Two-tailed tests were chosen in this study to test a null hypothesis of no effect of an item covariate, $H_0 : \gamma_d = 0$.

Paired-sample t test based on sum scores. A paired-sample t test can be used to test whether differences between the sum scores for subcategories of items defined on the basis of the item covariate are significantly different from 0. For example, an item covariate has values (0 and 1) and they correspond to two conditions, with the first five items in Condition 1 (covariate value of 0) and the remaining items in Condition 2 (covariate value of 1). With this example, the difference score can be calculated for each person p , that is, $d_p = x_{p2} - x_{p1}$, where $x_{p1} = \sum_1^5 y_{pi}$ and $x_{p2} = \sum_6^{10} y_{pi}$. The test statistic is as follows:

$$t = \frac{\bar{d} - 0}{\frac{S_d}{\sqrt{P}}}, df = P - 1, \quad (5)$$

where $\bar{d}$ and S_d are the mean and standard deviation of the sample difference scores, respectively, and P is the number of persons. Note that apart from the item covariate, both the LLTM and the LLTM-R assume item exchangeability, just as when sum scores are used in a repeated-measures design.

Profile Likelihood CI

CI approaches are in the first place of interest in an effect size context but they can also be used indirectly to test the null hypothesis of no effect of an item covariate, $H_0 : \gamma_d = 0$, by observing

whether 0 is within the estimated CI. Correspondence between null hypothesis testing (based on the z -statistic) and CI estimation is expected because z is a pivotal statistic whose distribution does not depend on the unknown population parameter. Because the z test was considered in this study, the Wald CI was not considered further. Instead the profile likelihood CI (Cox & Hinkley, 1974) was chosen. The profile likelihood CI allows for a nonquadratic log-likelihood. Although based on the asymptotic chi-square distribution of the LRT statistic, the profile likelihood CI does not assume normality of the estimator. An example of the profile likelihood CI calculation was provided in the Online Appendix.

The $100(1 - \alpha)\%$ profile likelihood CI for γ_d is the set of points where the profile likelihood function—that is, $\log L(\gamma_d)$ in the Online Appendix—exceeds the cut-off of $\log L(\hat{\gamma}_d, \hat{\boldsymbol{\delta}}) - \chi^2_{1-\alpha}(1)/2$, where $\boldsymbol{\delta}$ is a vector of the remaining parameters ($\boldsymbol{\delta} = [\mu, \sigma_\theta^2, \sigma_\beta^2]'$). The value of $\log L(\hat{\gamma}_d, \hat{\boldsymbol{\delta}})$ is fixed and obtained with log-likelihood function in R after the full model (the LLTM or LLTM-R with a covariate) is fit. Because of this definition one can expect the profile likelihood CI to yield very similar results as the LRT. However, the more extensive model to which the LRT is applied in comparison with the constrained model is a model in which all parameters are freely estimated, whereas for the profile likelihood CI one works with a sequence of fixed γ values in order to find those values that correspond with the selected $(1 - \alpha)$.

Model Comparisons

To test γ_d , the *null model* (i.e., the LLTM(-R) without an item covariate) can be compared with a *comparison model* (i.e., the LLTM(-R)) with the item covariate in question). The LRT and model information criteria, AIC (Akaike, 1974) and the BIC (Schwarz, 1978), can be used for the comparison.

LRT. LRT is for comparisons of the two nested models. Assume that the null hypothesis is given by $H_0 : \gamma_d = 0$, for some subspace $\Theta_{\gamma,0}$ of the parameter space Θ_γ of the fixed effect.

The LRT statistic is defined as

$$-2\log\lambda = -2\log \left[\frac{L(\hat{\gamma}_0)}{L(\hat{\gamma})} \right], \quad (6)$$

where L is the likelihood function, $\hat{\gamma}_0$ is the maximum likelihood estimate (MLE) obtained from maximizing L over $\Theta_{\gamma,0}$, and $\hat{\gamma}$ is the MLE obtained from maximizing L over Θ_γ . Asymptotically under $H_0 : \gamma_d = 0$, the LRT statistic follows a chi-square distribution with df equal to the difference between the two nested models (e.g., Casella & Berger, 2002).

Information criteria. The AIC is an estimate of the expected relative Kullback–Leibler (K-L) divergence (Akaike, 1974). Thus, the AIC implicitly estimates the divergence between the true model and the candidate model. Even though the actual K-L divergence is unknown because of the unknown true model, it was shown that the candidate model with the lowest AIC has the lowest expected K-L divergence (e.g., Burnham & Anderson, 2002). The AIC is efficient but it is not strongly consistent unless the true model is among the candidate models. The AIC penalizes for the number of parameters as follows:

$$\text{AIC} = -2\log L + 2 \times \text{Num}, \quad (7)$$

where *Num* is the number of parameters.

Schwarz (1978) derived the BIC to serve as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. The BIC is consistent, which means that it selects the true model as N tends to infinity (when the true model is among the candidate models or the number of parameters in the true model is finite) (see Claeskens & Hjort, 2008, Chapter 4, for details). The lowest BIC value is taken to indicate the best-fitting model. The BIC penalizes for the number of parameters (Num) more the larger the sample size is (N):

$$BIC = -2\log L + \log(N) \times Num. \quad (8)$$

In this study, the number of item responses ($P \times I$) was used for N as calculated in lme4 for item response models (Bates, 2005).

Burnham and Anderson (2002, pp. 298-301) compared the performance of the AIC and the BIC in selecting covariates in the linear regression model ($P = 253$) via a simulation study. They found that the AIC performed better than the BIC in terms of predictive mean square error. Vrieze (2012) found in a simulation study on factor models that the BIC ignored a small factor (with a small effect size of the loadings), unless the sample size is very large (e.g., $> 5,000$), whereas the AIC was indeed able to detect such factors regardless of sample size. Although the findings from Burnham and Anderson (2002, pp. 298-301) and Vrieze (2012) are obtained with other models than the LLTM(-R), there do not seem reasons to restrict the findings to those specific models. The findings are of a kind that suggests that they can be generalized.

The detection of item covariate effects using the methods described above was illustrated using an empirical dataset in the Online Appendix. The empirical dataset is one with multiple item covariates, in contrast with the simulation study that will be described next.

Simulation Study

A simulation study was designed to investigate the Type I error rate and the power for detecting an item covariate effect in various designs that may influence power and precision for the effects. The data generating models were the LLTM and the LLTM-R. The LLTM is a special case of the LLTM-R when the residual variance is 0. The same generated datasets were fit to the LLTM and the LLTM-R to investigate (a) misfitting, ignoring random residuals and thus fitting the LLTM when there is indeed residual variance, and (b) overfitting and thus fitting the LLTM-R when there is no residual variance.

Simulation Design

The following factors were used in the design of the simulation study: (a) number of items, (b) number of persons, (c) size of the item covariate effect γ , (d) the residual variance (across items), and (e) the nominal significance level (α). In this study, nominal significance levels, $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$, were considered. Previous research showed that the accuracy of the item effect estimates is affected by item covariate misspecification and collinearity among the item covariates (Baker, 1993; Green & Smith, 1987). To avoid such confounding effects, a correctly specified covariate was considered and just one as explained earlier. The covariate was effect coded and evenly distributed (e.g., -0.5 for five items and 0.5 for the other five items in the case of 10 items), in line with a balanced design as discussed earlier.

The number of items. The number of items was selected as $I = 10, 30$, and 80 . The number of 10 was considered small in item response theory (IRT) applications and was selected to investigate the use of an item covariate in a short test. The number of 30 was chosen to mimic the number of items in previous LLTM and LLTM-R applications (e.g., Embretson & Wetzel,

1987; Fischer, 1973; Gorin, 2005; Medina Díaz, 1993; Whitely & Schneider, 1981; see Table A1 in the Online Appendix). Weirich et al. (2014) found that the power of the LRT using the LLTM-R reached 0.94 when there are 80 items (with balanced incomplete block designs in NAEP) for a medium size effect (i.e., of the item position). Thus, the number of 80 was selected to examine the case with a large number of items.

The number of persons. The number of persons was set to $P = 200, 2,000$, and $4,000$. The number of 200 was chosen to approximate sample sizes in LLTM applications (e.g., Freund et al., 2008; Gorin, 2005; Kubinger, 2009; Medina Díaz, 1993; Whitely & Schneider, 1981; see Table A1 in the Online Appendix). Based on the findings in Weirich et al. (2014), the number of 4,000 was considered to examine the large sample properties of the investigated methods. A number of persons, 2,000, was chosen as a moderate level of sample size. This condition was also considered in Weirich et al. (2014) for the comparison with a level of 4,000.

Magnitude of covariate effect. Three levels of magnitudes (γ_d) were considered: 0, 0.2, and 0.5 as effects to be multiplied with an effect coded item covariate so that the differences are 0 (no effect), -0.2 versus 0.2 , and -0.5 versus 0.5 . To evaluate the impact of such an effect, the size of the residual variance is also important.

Residual variances. The residual variances were set to $\sigma_\epsilon^2 = 0, 0.2$, and 0.4 . Zero residual variance implies the LLTM as the true model, and with a residual variance of 0.2 and 0.4 , the LLTM-R is the true model. When item covariates are used to explain variability in item difficulty with the LLTM-R, it was found that the residual variance is less than 0.4 in papers that were surveyed. The values of 0.2 and 0.4 were chosen from the empirical studies in De Boeck (2008) and Hartig et al. (2012), respectively.

Combining the magnitude of the item covariate effect with the residual variance yields different values of the variance between items as well as different proportions of these variances explained by the item covariate. For a zero effect, the variances are $0, 0.2$, and 0.4 , depending on the three values of the residual variance ($0, 0.2, 0.4$), and of course, the proportion explained variance is either undefined ($0/0$) or zero ($0/0.2$ and $0/0.4$). For an effect of 0.2 , 0.04 needed to be added to the residual variance so that the variances are $0.04 (= 0 + 0.04)$, $0.24 (= 0.2 + 0.04)$, and $0.44 (= 0.4 + 0.04)$ and the explained proportions are $1.000, 0.167$, and 0.091 (correlations of $1.000, .408$, and $.302$), respectively. For an effect of 0.5 , 0.25 is need to be added to the residual variance so that the variances are $0.25 (= 0 + 0.25)$, $0.45 (= 0.2 + 0.25)$, $0.65 (= 0.4 + 0.25)$ and the explained proportions are $1.000, 0.556$, and 0.385 (correlations $1.000, .745$, and $.602$), respectively. The terms *small* and *large* were used for the two nonzero magnitudes of the covariate effect (0.2 and 0.5).

The person trait was generated with a standard normal distribution in each replication, $\theta_p \sim N(0, 1)$. Similarly, the item residuals were generated with a normal distribution in each replication (i.e., $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ for the cases of $\sigma_\epsilon^2 = 0.2, 0.4$, respectively). The average item difficulty parameter, μ , was set to 0 .

The four simulation conditions were fully crossed, yielding $81 (= 3 \times 3 \times 3 \times 3)$ conditions. About 1,000 replications were simulated for each of the 81 conditions.¹ Each generated dataset was analyzed using four models: an LLTM null model, an LLTM comparison model, an LLTM-R null model, and an LLTM-R comparison model.

Evaluation Measures

The proportion of 1,000 replications leading to an inference that γ_d is different from 0 was calculated. For the statistical tests, $\alpha = .01$, $\alpha = .05$, and $\alpha = .10$ were used. For the information

criteria (i.e., AIC, and BIC), strictly speaking, the notions of Type I error and power do not apply. Instead, the proportion that the estimation model with a nonzero effect (i.e., the LLTM in the case of $\sigma_\varepsilon = 0$; the LLTM-R in the case of $\sigma_\varepsilon^2 = 0.2$ or 0.4) is selected was calculated out of 1,000 replications. The proportion refers to false positives in the case of $\gamma_d = 0$ and to true positives in the case of $\gamma_d = 0.2$ and $\gamma_d = 0.5$. These proportions are the conceptual analogues of the Type I error rate and power, respectively. Ideally, the values for the Type I error rate meet the nominal significance level of α , and the values for power should be close to 1. As a rule of thumb, power of .90 and higher was considered satisfactory.

Hypotheses

The comparisons to be planned are more than exploratory. Expectations are based on the characteristic of the approaches. Therefore, the simulation study can be considered as being partly a hypothesis testing study with potentially important results for inferential practices. If the hypotheses turn out to be supported by the results, current practices may have to change. The study may suggest that some inferential approaches should be preferred on other approaches or that one model is a better basis to test covariate effects than another model.

Type I error rate. To formulate expectations regarding the Type I error rate, two categories of methods were differentiated: statistical tests and information criteria. The first category consists of the paired-sample t test with sum scores, the z test of the estimated effect, and the LRT for the comparison of two models. Although the profile likelihood CI is primarily meant as an effect size approach, it can also be used for null hypothesis testing. The second category consists of the AIC and the BIC.

In principle, the Type I error rate should be maintained at the nominal level. However, for some parts of the design, an inflated Type I error rate was expected for the paired-sample t test and the z test. For the paired-sample t test, the basis of the expectation is that when the item variance is larger than zero, the means as a function of the item covariate will be randomly different with a standard deviation of $\sigma_\varepsilon/\sqrt{I}$, where I is the number of items. When this random difference is tested based on a standard error (of the estimated effect) that is a function of the sample size P , the probability of rejecting the null hypothesis will increase and ultimately even reach 1.00 for an infinitely large sample size. The reason is that the sum score does not reflect any item variation. Therefore, it is expected that the Type I error inflation is larger as the number of items and the number of persons increase. The proper way to deal with this problem is of course to use the items as a random factor, unlike when sum scores are used.

Also for the z test, an inflated Type I error rate is expected when the number of items is small, because the z test assumes a normal (and thus a narrower) distribution of the estimated effect, which is not realized for smaller numbers of items in the presence of residual variance. This is a problem for the LLTM-R and more so the smaller the number of items is. However, because the z test does not ignore the item source of variance while the paired-sample t test does, the inflation is by far not expected to be as large for the former as for the latter. Because the LRT and the profile likelihood CI rely on the likelihood of the whole model and the number of df is clear (difference in number of parameters), it is expected that these approaches respect the nominal α -level better than the z test does.

The information criterion logic does not allow us to formulate an α -level, so that strictly speaking the nominal level cannot be a criterion to evaluate the results. However, because the AIC's penalty for model complexity is smaller than the BIC's penalty, the AIC tends to favor more complex models than the BIC. Therefore, the Type I error rate (strictly speaking the proportion of false positives) was expected to be higher for the AIC than for the BIC because the

model with a covariate effect is more complex than a model without. In addition, the Type I error rate of LRT is expected to be smaller than that of AIC and to be higher than that of BIC.²

Power. Power is expected to be based on two factors: the amount of information available in the data and the effect size. Therefore, it is expected that the power increases as the number of persons and the number of items increase, and that the power increases with the magnitude of the effect and decreases with the residual variance. This is because the explained variance (and thus the standardized effect size calculated in terms of correlations) increases with the magnitude of the item covariate effect and decreases with the residual variance. Furthermore, because of the same problems as with the Type I error inflation, it is expected that the paired-sample t test has a much larger power and the z test has a slightly larger power compared with the LRT and the profile likelihood CI for the case the residual variance is larger than zero. Based on the same reasons as for the Type I error rate, the AIC is expected to have a higher power (strictly speaking a higher proportion of true positives) than the BIC, and the power of LRT is expected to be smaller than that of AIC and to be higher than that of BIC.

Using a different model than the data generation model. When the data-estimation model is different from the data generation model, the two estimation models were compared for the same data: the LLTM and the LLTM-R were both used as estimation models for the LLTM and the LLTM-R as data generation models. Using the LLTM for LLTM-R generated data means that a clearly misfitting (too constrained) model is used because the true residual variance is ignored. Using the LLTM-R for LLTM generated data means that an overfitting model (not sufficiently constrained) is used by allowing for a nonexistent true residual variance. For the LLTM-R used for LLTM data a somewhat deflated Type I error rate is expected because the estimate of the residual variance, even when very small, will contribute to the uncertainty of the item covariate estimated effect. For the LLTM used for LLTM-R data, an inflated Type I error rate and an overestimated power are expected because the uncertainty of the residual variance is not taken into account.

The hypotheses for the Type I error rate and power are highly related. This is of course because the power depends on the Type I error rate. Still it is possible to have an adequate Type I error rate, one that corresponds with the nominal α , in combination with a high or low power rate. And, depending on the actual distribution of the test statistic (when the distribution deviates from the theoretical one), it is even possible that when two methods are compared, the Type I error rate of one is smaller (or larger) and the power is higher (or lower), although that is not the common case. Generally speaking, if the Type I error rate is higher (or lower), one can expect a higher (or lower) power, one that cannot be interpreted as an advantage of the method but rather as the consequence of a more liberal (or conservative) testing.

Results

The results are presented in Figure 1 for the Type I error rate, in Figure 2 for power, in Figure 3 for the Type I error rate and power of a misfitting model (using the LLTM for LLTM-R data), and in Figure 4 for the Type I error and power of an overfitting model (using the LLTM-R for LLTM data). For the comparison purposes, results in the case of $\sigma_{\epsilon}^2 = 0$ (presented in Figures 1 and 2) were also presented in Figure 3 and results in the case of $\sigma_{\epsilon}^2 = 0.2$ or 0.4 (presented in Figures 1 and 2) were also presented in Figure 4. Only the results for $\alpha = .05$ are presented in the figures. The conclusions for $\alpha = .01$ and $\alpha = .10$ are the same. Results of the profile likelihood CI were not presented in the figures because the results were similar to those of LRT. For the numerical results including the profile likelihood CI for all α levels, see Tables A2 to A6 in the Online Appendix.

Figure 1. Type I error rate (for $\gamma=0$) in the case of $\sigma_\epsilon^2=0$ (top) and in the case of $\sigma_\epsilon^2=0.2$ and $\sigma_\epsilon^2=0.4$ (bottom).

Note. The nominal .05 level is indicated with a straight horizontal line. t = paired-sample t test; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Figure 2. Power (for $\gamma=0.2$ and $\gamma=0.5$) in the case of $\sigma_e^2=0$ (top) and in the case of $\sigma_e^2=0.2$ and $\sigma_e^2=0.4$ (bottom).
Note. The satisfactory .90 level is indicated with a straight horizontal line. t = paired-sample t test; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion.

Figure 3. Type I error rate (for $\gamma=0$) (top) and power (for $\gamma=0.2$ and $\gamma=0.5$) (bottom) for an overfitting model in the case of $\sigma^2_\epsilon = 0$ (LLTM-R results).
Note. The nominal .05 level (top) and the satisfactory .90 level (bottom) are indicated with a straight horizontal line. LLTM results are presented in gray line for comparison purposes. For the paired-sample t test, there is no difference between the LLTM and LLTM-R because the paired-sample t test depends only on the data generation model. The nominal .05 level for the Type I error is indicated in the top part of the figure with a straight horizontal line. t = paired-sample t test; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion; LLTM = linear logistic test model; LLTM-R = LLTM with random item residuals.

Figure 4. Type I error rate (for $\gamma=0$) (top) and power (for $\gamma=0.2$ and $\gamma=0.5$) (bottom) under misfit in the case of $\sigma_{\epsilon}^2=0.2$ and $\sigma_{\epsilon}^2=0.4$. (LLTM results).

Note. The nominal .05 level (top) and the satisfactory .90 level (bottom) are indicated with a straight horizontal line. LLTM-R results are presented in gray line for comparison purposes. For the paired-sample t test, there is no difference between the LLTM and LLTM-R because the paired-sample t test depends only on the data generation model. The nominal .05 level for the Type I error is indicated in the top part of the figure with a straight horizontal line. t = paired-sample t test; LRT = likelihood ratio test; AIC = Akaike information criterion; BIC = Bayesian information criterion; LLTM = linear logistic test model; LLTM-R = LLTM with random item residuals.

Type I error rate. Type I error rate as presented in the top half of Figure 1(top) was calculated for the case of $\sigma_{\varepsilon}^2 = 0$ using the LLTM for LLTM data, whereas the bottom half presents the Type I error rate for the case of $\sigma_{\varepsilon}^2 = 0.2$ and $\sigma_{\varepsilon}^2 = 0.4$, using the LLTM-R for LLTM-R data.

For the case of $\sigma_{\varepsilon}^2 = 0$ the Type I error rate was roughly equal to the nominal level for the paired-sample t test, the z test, profile likelihood CI, and LRT. However, in line with expectations, the Type I error rate was larger for the AIC than for the BIC, and higher for the AIC than for the statistical tests, while the Type I error rate of the BIC was lower than for the statistical tests.

The following patterns were observed for $\sigma_{\varepsilon}^2 = 0.2$ and $\sigma_{\varepsilon}^2 = 0.4$. First, as expected, the Type I error rates for the paired-sample t test were highly inflated, and more so for the larger sample sizes and the larger σ_{ε}^2 . Second, and again as expected, a slightly inflated Type I error rate was found for the z test, to a lesser extent when the number of items is large. Also for the profile likelihood CI and LRT, an inflated Type I error rate was found, although smaller than for the z test, again as expected. As one can tell from the results, the overinflation has disappeared and the nominal α level is perfectly realized for the profile likelihood CI and the LRT and for the z test when 80 items are used. Third, and again as expected, the Type I error rate was larger for the AIC than for the statistical tests and higher for the latter than for the BIC.

Power. Power as presented in the top half of Figure 2 was calculated for the case of $\sigma_{\varepsilon}^2 = 0$ using the LLTM for LLTM data, whereas the bottom half of Figure 2 presents power for the case of $\sigma_{\varepsilon}^2 = 0.2$ and $\sigma_{\varepsilon}^2 = 0.4$, using the LLTM-R for LLTM-R data. In the case of $\sigma_{\varepsilon}^2 = 0$, the targeted power of .90 was always reached for $\gamma = 0.5$. For $\gamma = 0.2$, the targeted power of .90 was reached when either the number of items was larger than 10 or the number of persons was larger than 200. Both were sufficient conditions, except for the BIC when the number of items was 30. Because a perfect power level was almost always reached, it was difficult to notice the effect of γ except for a small number of items and persons.

The following trends showed in the case of $\sigma_{\varepsilon}^2 = 0.2$ and $\sigma_{\varepsilon}^2 = 0.4$. First, for all statistical tests and information criteria, the power increased for larger numbers of items, larger numbers of persons, larger item covariate effects, and smaller residual variances. This is an evident result in line with the principles of power as explained earlier. Second, as expected, the power of the paired-sample t test was higher than the power of other tests, followed by the z test. Third, for most methods, the level of .90 was not reached with 10 items. With 30 items, the desired level of .90 was also reached with the AIC and not only with the paired-sample t test and it was almost reached with the z test and the LRT (and profile likelihood CI) if the item covariate effect was 0.5 and the residual variance was 0.2. For a covariate effect of 0.2, 80 items and 4,000 persons did not suffice to reach a power level of 0.90. Fourth, as expected, the power level of the AIC is always higher than the power level of the BIC. Although the AIC has higher power levels than most tests, the BIC has always lower power levels than most tests. For the BIC to reach the targeted power level, the number of items needed to be $I = 80$ when $\gamma = 0.5$.

Using a different model than the data generation model. When the LLTM-R was used in the absence of residual variance (see Figure 3), the Type I error rate was slightly deflated and the power was also slightly lower than when the LLTM was used. The differences showed primarily for the small sample size ($P = 200$). These results confirm the result hypotheses. The LLTM-R adds unnecessary uncertainty in the absence of residual variance.

The results are more drastic when the LLTM was used in the case of $\sigma_{\varepsilon}^2 = 0.2$ and $\sigma_{\varepsilon}^2 = 0.4$ (see Figure 4). As expected, the Type I error rate was inflated for all conditions, but the inflation was extremely large and applies across all methods and for all conditions in the simulation design. These results confirm the high inflation rates when the paired-sample t test is used. The common source is that the item variance is ignored. In addition, as also expected, the power

was much higher using the LLTM in the presence of residual variance compared with using the LLTM-R. Also these results follow from ignoring the uncertainty due to the residual variance and they can therefore not be seen as a positive quality of the LLTM.

Summary and Discussion

The LLTM and LLTM-R have been used to investigate the effects of item covariates (e.g., cognitive variables, item characteristics, item design factors). In this study, the paired-sample *t* test based on sum scores, the *z* test, profile likelihood CI, LRT, AIC, and BIC were evaluated via Monte Carlo studies. Based on the simulation results, the following conclusions can be drawn. First, the LRT seems a better test than the *z* test. It does respect the nominal α level better, although a large set of items is needed to reach the nominal level. Although the profile likelihood CI is expected to perform better for small sample size, its performance is highly similar to that of the LRT.

Second, the information criteria are not a good basis for inference regarding the effect of a covariate. The AIC leads to a large proportion of false positives, and the BIC results in a large proportion of false negatives. These findings are in line with the common knowledge that the AIC favors more complex models and that the BIC is conservative if the sample size is not very large (e.g., Burnham & Anderson, 2002). Although the number of persons was very large in one of the simulation conditions ($P=4,000$), the number of items is still small even when 80 items may seem large for a test.

Third, the LLTM-R is clearly the better choice as an estimation model. In the presence of residual variance, the LLTM leads to largely invalid tests and an extremely high proportion of false positives. The better approach is to start with the LLTM-R to check whether there is residual variance. Only when there is no indication at all for residual variance should the LLTM be used. Note that even when the correlation between the item covariate and the item parameters is 0.745 ($\gamma=0.5$ and a residual variance of 0.2), the power is largely overestimated. Staying with the LLTM-R has no dramatic consequences, it would only lead to a somewhat conservative testing approach. Because the problematic consequences of the LLTM in the presence of residual variance increase with the number of persons, the LLTM is more problematic for large-scale studies than for smaller-scale studies. In line with the problematic features of the LLTM, if the item covariate-based explanation is not perfect, also the paired-sample *t* test based on sum scores is problematic. This is a nicely convergent finding, but it has less practical value because one would normally not consider the paired-sample *t* test in an IRT context. However, the result is relevant for more traditional approaches to repeated measures if sum scores are used.

Fourth, to reach the nominal Type I error rate and the desired power rate of .90, a very large number of items (e.g., >80), and for power additionally a large effect size is needed, unless the LLTM happens to be the true model, which is unlikely because it requires a perfect or almost perfect explanation of the item variance. Because very large numbers of items are practically impossible, one should either look for item covariates with an (almost) perfect explanatory value, or one should accept the suboptimal qualities of one's approach. The application in the Online Appendix shows that it pays off to add item covariates with a high explanatory value.

Not all these recommendations are equally strong. For example, although the profile likelihood CI and the LRT seem to do better than the *z* test, there is still room for improvement because they still show some Type I error inflation for smaller numbers of items. Regarding power, a major concern is that either the LLTM is needed, which implies a (nearly) perfect explanation of the item difficulties, or a really large number of items. It is worth considering

other possible purposes that can be fulfilled with the LLTM-R, other than null hypothesis testing of item covariate effects. For example, the LLTM-R can be used for item generation purposes (De Boeck, Cho, & Wilson, 2016) and the model may still be useful for the item part of a model when the model is used in the first place to derive ability scores and not so much to estimate item effects. Finally, the authors do not want to depreciate the AIC and BIC as general model selection methods. The information criteria were used in a similar way as null hypothesis testing methods, which they are not, as explained earlier.

The simulation conditions employed in the study are limited to one item covariate. When multiple item covariates are considered, one may hope that residual item variance is further reduced, so that the ideal situation is better approached. However, the residual variance as such seems important, and adding item covariates does not imply a high proportion of item variance can be explained. In case multiple item covariates are available, a new problem arises, which is the selection of the best model, with the best set of item covariates. The number of comparison models is $2^D - 1$, where D is the number of possible item covariates. It is easy to see that the number of possible comparison models rapidly increases as a function of D . Thus, to investigate the Type I error rate and power with multiple item covariates, the best-fitting model should be selected first. An evaluation of item covariate selection methods in the LLTM and LLTM-R is beyond the scope of this article, but a real data application with multiple item covariates was included in the Online Appendix.

Another limitation is the type of item covariate. In this study, a balanced item covariate with an equal number of items per value of the item covariate was used. One may expect that the power is affected in unbalanced designs. In addition, the item covariate was binary, but that does not seem to be a factor that affects power in the GLMM framework (e.g., Johnson et al., 2015). Finally, Green and Smith (1987) found that collinearity in continuous item covariates (e.g., frequency with which an item attribute d is needed) affected the accuracy of item covariate effect estimates in the LLTM, which is similar result as for linear regression. In the presence of collinearity, they found via a simulation study that the item covariate estimates and standard errors were not accurate when the sample size was small (less than 200) (see Table 3, Green & Smith, 1987, p. 378).

In sum, the authors believe that the results in the current study are informative even when limited to a balanced binary covariate. The results concern specific but basic issues regarding the LLTM and the LLTM-R, such as ignoring residual item variation. However, one may expect that factors that play a role in regular linear regression would also play a role in the LLTM and LLTM-R, such as model selection, collinearity, and heteroscedasticity of error terms. To evaluate the more precise effects of these factors, further studies with more complex simulation studies would be required.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplement Material

The online appendices are available at <http://journals.sagepub.com/doi/suppl/10.1177/0146621617692078>.

Notes

1. The number of 1,000 replications was chosen because it is the maximum number in the literature the accuracy of item response theory (IRT) estimates and standard errors were investigated (e.g., see Table 1 in Forero & Maydeu-Olivares, 2009).
2. For the model comparison between the null model and the comparison model (i.e., the nested model, $df = 1$), the critical value for the LRT is 3.84, $= \chi^2(0.95, df)$ with $\alpha = .05$, the model complexity term for the AIC differences between the two models is $2 \times df = 2$, and the model complexity term for the BIC differences between the two models is $\log(N) \times df = \log(200 \times 10) = 7.6$ for $J = 200$ and $I = 10$ as the smallest number of persons and items in the simulation condition. Thus, the following order relation is expected, Type I error rate(AIC) > Type I error rate(LRT) > Type I error rate(BIC). The authors thank an anonymous reviewer of a previous version of this article for pointing out this order relation among AIC, LRT, and BIC for the Type I error rate.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716-723.
- Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59, 390-412.
- Baker, F. B. (1993). Sensitivity of the linear logistic test model to misspecification of the weight matrix. *Applied Psychological Measurement*, 17, 201-210.
- Bates, D. (2005). *lme4: Mixed-effects modeling with R*. New York, NY: Springer.
- Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using Eigen and R syntax* [R package version 0.999375-39]. Retrieved from <https://cran.r-project.org/web/packages/lme4/index.html>
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York, NY: Springer.
- Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Wadsworth & Brooks.
- Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement*, 52, 200-222.
- Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, 55, 12-25.
- Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, UK: Cambridge University Press.
- Cox, A. Z., & Hinkley, A. Z. (1974). *Theoretical statistics*. London, England: Chapman & Hall.
- De Boeck, P. (2008). Random item IRT models. *Psychometrika*, 73, 533-559.
- De Boeck, P., Cho, S.-J., Wilson, M. (2016). Explanatory item response models: An approach to cognitive assessment. In A. Rupp & J. Leighton (Eds.), *Handbook of cognition and assessment* (pp. 249-266). Harvard, MA: Wiley Blackwell.
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.
- Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.
- Embretson, S. E., & Wetzels, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175-193.
- Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, 27, 359-374.
- Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: Limited versus full information methods. *Psychological Methods*, 14, 275-299.
- Freund, P. A., Hofer, S., & Holling, H. (2008). Explaining and controlling for the psychometric properties of computer-generated figural matrix items. *Applied Psychological Measurement*, 32, 195-210.

- Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, 42, 351-373.
- Green, K. E., & Smith, R. M. (1987). A comparison of two methods of decomposing item difficulties. *Journal of Educational Statistics*, 12, 369-381.
- Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement*, 72, 665-686.
- Hecht, M., Weirich, S., Siegle, T., & Frey, A. (2015). Effects of design properties on parameter estimation in large-scale assessments. *Educational and Psychological Measurement*, 75, 1021-1044.
- Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement*, 7, 732-746.
- Hornke, L. F., & Habon, M. W. (1986). Rule-based item bank construction and evaluation within the linear logistic framework. *Applied Psychological Measurement*, 10, 369-380.
- Janssen, R., Schepers, J., & Peres, D. (2004). Models with item and item group predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 189-212). New York, NY: Springer.
- Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, 52, 5066-5074.
- Johnson, P. C. D., Barry, S. J., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, 6, 133-142.
- Kelly, K., & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods*, 8, 305-321.
- Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement*, 69, 232-244.
- Medina Díaz, M. (1993). Analysis of cognitive structure using the linear logistic test model and quadratic assignment. *Applied Psychological Measurement*, 17, 117-130.
- Mislevy, R. J. (1988). Exploiting auxiliary information about items in the estimation of Rasch item difficulty parameters. *Applied Psychological Measurement*, 12, 281-296.
- Molenberghs, G., & Verbeke, G. (2004). *Models for discrete longitudinal data*. New York, NY: Springer.
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. New York, NY: Springer.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6, 461-464.
- Sheenan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, 27, 255-272.
- Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 169-193). New York, NY: Academic Press.
- Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164-182.
- Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, 17, 228-243.
- Wald, A., & Wolfowitz, J. (1939). Confidence limits for continuous distribution functions. *The Annals of Mathematical Statistics*, 10, 105-118.
- Weirich, S., Hecht, M., & Böhme, K. (2014). Modeling item position effects using generalized linear mixed models. *Applied Psychological Measurement*, 38, 535-548.
- Whitely, S. E., & Schneider, L. M. (1981). Information structure for geometric analogies: A test theory approach. *Applied Psychological Measurement*, 5, 383-397.