Derivation and Applicability of Asymptotic Results for Multiple Subtests Person-Fit Statistics

Casper J. Albers, Rob R. Meijer, Jorge N. Tendeiro

University of Groningen

Department Psychometrics and Statistics, Faculty of Behavioral and Social Sciences, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, The Netherlands. Corresponding author: c.j.albers@rug.nl, www.casperalbers.nl

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions in this report are those of the author(s) and do not necessarily reflect position or policy of LSAC.

Derivation and Applicability of Asymptotic Results for Multiple Subtests Person-Fit Statistics

Abstract

In high-stakes testing, it is important to check the validity of individual test scores. Although a test may, in general, result in valid test scores for most test takers, for some test takers test scores may not provide a good description of a test taker's proficiency level. Person-fit statistics have been proposed to check the validity of individual test scores. In this study we first discuss the theoretical asymptotic sampling distribution of two person-fit statistics that can be used for tests that consist of multiple subtests. Second, we conducted a simulation study to investigate the applicability of this asymptotic theory for tests of finite length, in which we varied the correlation between subtests and number of items in the subtests. We showed that these distributions provide reasonable approximations, even for tests consisting of subtests of only 10 items each. These results have practical value because researchers do not have to rely on extensive simulation studies to simulate sampling distributions.

Keywords: item response theory, person-model fit, validity test scores

Derivation and Applicability of Asymptotic Results for Multiple Subtests Person-Fit Statistics

In high-stakes testing, individual test scores are being used to make important decisions for individual test takers. In these circumstances it is important to check the validity of the individual test scores. Although test scores may be valid for most persons in a particular population, for some test takers these scores may not reflect their true proficiency level. For example, Tendeiro and Meijer (2014) showed that for some test takers on a high-stakes test the proficiency scores did not seem to reflect their true proficiency level. Several methods have been proposed to check the validity of individual test scores. In this study we focus on methods that are sensitive to the fit of individual response patterns to an item response theory (IRT) model. The idea behind this approach is that, when an item score pattern is very unexpected given the estimated proficiency level, this estimated proficiency level might not provide a good estimate of their true proficiency level. These methods are often denoted as person-fit methods or person-fit statistics (Meijer & Sijtsma, 2001).

One of the most popular statistics is the standardized log-likelihood statistic, denoted l_z, proposed by Drasgow, Levine, and Williams (1985). Based on asymptotic arguments, Snijders (2001; see also Magis, Raîche, & Béland, 2012) suggested an improved version of this statistic, denoting it l_z^*. In assessing person fit, the person parameter θ is unknown and needs to be estimated by $\hat{\theta}$. This estimation process biases the (asymptotic) behavior of l_z and Snijders' version accounts for this bias.

Both l_z and l_z^* are developed for unidimensional tests. In practice, however, many tests consist of several (correlated) subtests. For example, the Law School Admission Test consists of four subtests which total scores are combined into one total score. For these types of tests it would be useful to have a person-fit statistic that combines information from the multiple subtests into one person-fit value. Drasgow, Levine, and Williams (1991) proposed a multiple subtest extension of l_z, which they denoted l_{zm}. Conijn, Emons, and Sijtsma (2014) compared several approaches based on l_z to studying person-fit for noncognitive multiple subtests consisting of polytomous items. Because of the advantage of l_z^* over l_z, Tendeiro, Meijer, and Albers (2014) recently studied the performance of Conijn et al.'s (2014) approaches applied to l_z^* rather than l_z for multiple subtests settings based on dichotomous items. This study by Tendeiro et al. (2014) was performed on the basis of a simulation design. The aim of the current paper is to study the distributional properties of multi-subtest modifications of l_z^* statistic through statistical (asymptotic) theory.

The outline of this report is as follows. In the next section, we introduce the l_z (Drasgow et al., 1985) and l_z^* (Snijders, 2001) statistics. Next, multiple subtests extensions based on l_z (Drasgow et al., 1991; Conijn et al., 2014) will be discussed. As explained above, the l_z^* statistic is a bias-removing improvement upon the l_z statistic. In the main theoretical section of this paper we study the theoretical null distribution of the multiple subtests person-fit statistics based on l_z^* instead of on l_z. The distributional theory is based on asymptotic arguments. The length of each subtest and the correlation of the latent traits between subtests are manipulated by means of a simulation study. The

goal is to study possible effects of these factors on the quality of the asymptotic approximations. It will be shown that the asymptotic approximations are fairly good for subtest lengths as low as 10 items.

The l_z and l_z^* statistics

This section shortly describes the l_z and l_z^* statistics. For a more extensive discussion of the l_z statistic see, for example, Armstrong, Stoumbos, Kung and Shi (2007), Magis et al. (2012), and van Krimpen-Stoop and Meijer (1999).

The l_z statistic

A test taker with trait level θ is administered a univariate test consisting of n items. The random variable X_i equals 0 or 1, depending on whether item i was answered correctly or incorrectly, respectively. The probability of answering correctly, $P(X_i = 1 | \theta)$, is denoted by $p_i(\theta)$. The three-parameter logistic model (3PLM; see Embretson & Reise, 2000), or its constrained versions known as the two- and one-parameter logistic models, are commonly used in IRT to describe the stochastic relationship between θ and X_i. The 3PLM is given by

$$p_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}}, \quad (1)$$

where a_i, b_i, and c_i denote the discrimination, difficulty, and pseudo-guessing parameters of item i. The two-parameter logistic model (2PLM), which results from constraining c_i to zero in Equation 1, will be used in our simulation study. However, the theory in this paper applies to other models as well.

The likelihood function of a response vector $\mathbf{X} = (X_1, \ldots, X_n)$ is given by

$$L(\theta) = \prod_{i=1}^{n} p_i(\theta)^{X_i} \left(1 - p_i(\theta)\right)^{1-X_i},$$

and the maximum likelihood estimator $\hat{\theta}_{\mathrm{ML}}$ is obtained by maximizing $L(\theta)$ or, equivalently, by maximizing the log-likelihood

$$l_0(\theta) = \log\left(L(\theta)\right) = \sum_{i=1}^{n} \left\{ X_i \log\left(p_i(\theta)\right) + (1 - X_i) \log\left(1 - p_i(\theta)\right) \right\}. \qquad (2)$$

Since this function depends on the number of items, it is not directly applicable as a person-fit statistic. To this end, Drasgow et al. (1985) proposed to use the standardized version

$$l_z = \frac{l_0 - E(l_0)}{\sqrt{Var(l_0)}}$$

as a person-fit statistic. Here, the expectation and variance are given by

$$E(l_0) = \sum_{i=1}^{n} \left\{ p_i(\theta) \log\left(p_i(\theta)\right) + \left(1 - p_i(\theta)\right) \log\left(1 - p_i(\theta)\right) \right\} \qquad (3)$$

and

$$Var(l_0) = \sum_{i=1}^{n} \left\{ p_i(\theta)\left(1 - p_i(\theta)\right) \left[\log\left(p_i(\theta)\right) - \log\left(1 - p_i(\theta)\right) \right]^2 \right\}, \qquad (4)$$

respectively. For known trait parameters θ and under the assumption of local independence, l_z is asymptotically standard normal, where "asymptotically" refers to the length n of the test. Throughout the paper, item parameters are assumed known.

The l_z^* statistic

Snijders (2001) argued that, in practice, it is rarely the case that true trait values θ are known. He showed that the l_z statistic is biased when the true θ is replaced by an estimate $\hat{\theta}$. He proposed a correction, actually applicable to a wider range of estimators

than l_z. Snijders (2001) studied the class of standardized person-fit statistics described through

$$\frac{W_n(\theta)}{\sqrt{Var(W_n(\theta))}}, \tag{5}$$

where $W_n(\theta) = \sum_{i=1}^{n}(X_i - p_i(\theta))w_i(\theta)$, with $w_i(\theta)$ a particular choice of a weight function. For $w_i(\theta) = \log(p_i(\theta)) - \log(1 - p_i(\theta))$, the l_z statistic is obtained.

Snijders (2001) showed that the bias introduced by replacing θ by its estimate $\hat{\theta}$ does not vanish asymptotically, causing, for example, conservative inferences in case of parameter estimation through the 3PLM. A solution to this problem is obtained by modifying the weights $w_i(\theta)$ via

$$\tilde{w}_i(\theta) = w_i(\theta) - c_n(\theta) r_i(\theta),$$

where

$$c_n(\theta) = \frac{\sum_{i=1}^{n} p_i'(\theta) w_i(\theta)}{\sum_{i=1}^{n} p_i'(\theta) r_i(\theta)}, \tag{6}$$

$p_i'(\theta)$ is the first-order derivative of $p_i(\theta)$ with respect to θ, and the $r_i(\theta)$ are chosen such that

$$r_0(\hat{\theta}) + \sum_{i=1}^{n}(X_i - p_i(\hat{\theta}))r_i(\hat{\theta}) = 0. \tag{7}$$

Various choices of $r_i(\hat{\theta})$ satisfy this relation. The most common choice is that of the maximum likelihood estimates, given by $r_0(\hat{\theta}) = 0$ and $r_i(\hat{\theta}) = \frac{p_i'(\hat{\theta})}{p_i'(\hat{\theta})(1-p_i'(\hat{\theta}))}$ ($i > 0$).

Snijders (2001) showed that $W_n\left(\hat{\theta}\right)$ is asymptotically normally distributed with expected value

$$E\left(W_n\left(\hat{\theta}\right)\right) = -c_n\left(\hat{\theta}\right) r_0\left(\hat{\theta}\right)$$

and variance $Var\left(W_n\left(\hat{\theta}\right)\right) = n\tau_n^2\left(\hat{\theta}\right)$, with

$$\tau_n^2\left(\hat{\theta}\right) = \frac{1}{n}\sum_{i=1}^{n} \tilde{w}_i^2\left(\hat{\theta}\right) p_i\left(\hat{\theta}\right)\left(1 - p_i\left(\hat{\theta}\right)\right). \tag{8}$$

As a consequence, the statistic

$$l_z^* = \frac{W_n\left(\hat{\theta}\right) - E\left(W_n\left(\hat{\theta}\right)\right)}{\sqrt{Var\left(W_n\left(\hat{\theta}\right)\right)}}$$

is asymptotically standard normally distributed.

In case the estimation of θ is done through the method of maximum likelihood (ML) things become slightly easier. In this case $r_0\left(\hat{\theta}\right) = 0$, which implies that $\mathrm{E}\left(W_n\left(\hat{\theta}\right)\right) = 0$. In this paper we will only work with ML estimators, but we shall continue to use the generalized notation of Snijders (2001).

The $l_{zm}^{(\mathrm{D})}$ and $l_{zm}^{(\mathrm{C})}$ statistics and proposed corrections

The multiple subtests statistic developed by Drasgow et al. (1991), here denoted $l_{zm}^{(\mathrm{D})}$, is based on the sum of the l_0 statistics for each subtest. For each subtest *s* (*s* = 1, …, *S*) one computes $l_{0(s)}$, $E\left(l_{0(s)}\right)$, and $Var\left(l_{0(s)}\right)$ as described above. Then,

$$l_{zm}^{(\mathrm{D})} = \frac{\left(\sum_{s=1}^{S} l_{0(s)}\right) - \left(\sum_{s=1}^{S} E\left(l_{0(s)}\right)\right)}{\left[\sum_{s=1}^{S} Var\left(l_{0(s)}\right)\right]^{1/2}}. \tag{9}$$

The lack of covariances in the denominator is a consequence of the IRT assumption of local independence. Note that this assumption implies that the $l_{0(s)}$ scores across subtests are independent, but it still allows for the latent traits $\theta_{i(s)}$ across subtests to be correlated, which often is the case in practice. Assuming that the true scores θ are known, $l_{zm}^{(\mathrm{D})}$ is standard normally distributed.

Conijn et al. (2014) suggested a slightly different approach to compute the multiple subtests statistic. Rather than summing the $l_{0(s)}$ statistics over the S subtests and then standardizing the sum, Conijn et al. suggested to first standardize each $l_{0(s)}$ and then to sum the standardized $l_{z(s)}$ statistics:

$$l_{zm}^{(\mathrm{C})} = \sum_{s=1}^{S} l_{z(s)}.$$

This approach is based on the same assumptions as the approach by Drasgow et al. (1991). In our simulation study we shall study which method performs better.

Just as the l_z approach is biased when θ is unknown, so are the multiple subtest extensions $l_{zm}^{(\mathrm{D})}$ and $l_{zm}^{(\mathrm{C})}$. The solution by Snijders (2001) is actually directly applicable to $l_{zm}^{(\mathrm{D})}$ and $l_{zm}^{(\mathrm{C})}$. We therefore propose two new person-fit statistics, which we denote by $l_{zm}^{*(\mathrm{D})}$ and $l_{zm}^{*(\mathrm{C})}$, respectively. In the next section, the asymptotic null distribution of each of these statistics is derived.

Asymptotic null distribution of $l_{zm}^{*(\mathrm{D})}$ and $l_{zm}^{*(\mathrm{C})}$

In this section we derive the asymptotic distributions of $l_{zm}^{*(\mathrm{D})}$ and $l_{zm}^{*(\mathrm{C})}$. In the next section we study the applicability of this asymptotic theory for tests of finite length.

Null distribution of $l_{zm}^{(\mathrm{D})}$ and $l_{zm}^{*(\mathrm{D})}$

Statistic $l_{zm}^{(\mathrm{D})}$ is asymptotically normally distributed if the true θ values are used (Drasgow et al., 1991). However, replacing true with estimated θ values introduces a bias, as explained above. We shall now try to correct this bias by applying Snijder's approach.

It is possible to rewrite $l_{zm}^{(\mathrm{D})}$ given by Equation 9 in the form of Equation 5:

$$l_{zm}^{(\mathrm{D})} = \frac{W(\theta)}{\sqrt{Var(W(\theta))}}, \tag{10}$$

with

$$W(\theta) = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\left[X_{i(s)} - p_{i(s)}(\theta_s)\right] w_{i(s)}(\theta_s), \tag{11}$$

$$w_{i(s)}(\theta_s) = \log\left(p_{i(s)}(\theta_s)\right) - \log\left(1 - p_{i(s)}(\theta_s)\right).$$

Here, θ denotes the vector (θ_1, …, θ_S) of latent trait parameters per subtest. That Equation 9 can be written as Equation 10 can be seen as follows. First, for the numerator (recall Equations 2 and 3),

$$\sum_{s=1}^{S} l_{0(s)} = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\left\{X_{i(s)} \log\left(p_{i(s)}(\theta_s)\right) + \left(1 - X_{i(s)}\right)\log\left(1 - p_{i(s)}(\theta_s)\right)\right\}$$

and

$$\sum_{s=1}^{S} E\left(l_{0(s)}\right) = \sum_{s=1}^{S}\sum_{i=1}^{n_s}\left\{p_{i(s)} \log\left(p_{i(s)}(\theta_s)\right) + \left(1 - p_{i(s)}\right)\log\left(1 - p_{i(s)}(\theta_s)\right)\right\},$$

thus $\sum_{s=1}^{S} l_{0(s)} - \sum_{s=1}^{S} E\left(l_{0(s)}\right) = W(\theta)$. For the denominator, we have (recall Equation 4)

$$\sum_{s=1}^{S} Var\left(l_{0(s)}\right) = \sum_{s=1}^{S}\sum_{i=1}^{n_s} p_{i(s)}\left(\theta_s\right)\left(1-p_{i(s)}\left(\theta_s\right)\right)\left[\log\left(p_{i(s)}\left(\theta_s\right)\right) - \log\left(1-p_{i(s)}\left(\theta_s\right)\right)\right]^2$$

and

$$Var\left(W(\theta)\right) = \sum_{s=1}^{S}\sum_{i=1}^{n_s} p_{i(s)}\left(\theta_s\right)\left(1-p_{i(s)}\left(\theta_s\right)\right)\left(w_{i(s)}\left(\theta_s\right)\right)^2,$$

and therefore $\sum_{s=1}^{S} Var\left(l_{0(s)}\right) = Var\left(W(\theta)\right)$.

We have established that $l_{zm}^{(\mathrm{D})}$ belongs to the family of statistics considered by Snijders. It therefore follows that, for trait estimates $\hat{\theta}_{(s)}$ satisfying Equation 7, $W(\hat{\theta})$ is asymptotically normally distributed with expected value

$$E\left(W\left(\hat{\theta}\right)\right) = \sum_{s=1}^{S} -c_{n_s}\left(\hat{\theta}_{(s)}\right) r_{0(s)}\left(\hat{\theta}_{(s)}\right)$$

and variance

$$Var\left(W\left(\hat{\theta}\right)\right) = \sum_{s=1}^{S} n_s \tau_{n_s}^2\left(\hat{\theta}_{(s)}\right),$$

where $c_{n_s}\left(\hat{\theta}_{(s)}\right)$ and $\tau_{n_s}^2\left(\hat{\theta}_{(s)}\right)$ are the functions defined by Equations 6 and 8 applied to subtest *s*.

Thus, in the above we established that, under the assumption of local independence, asymptotically

$$l_{zm}^{*(\mathrm{D})} = \frac{W\left(\hat{\theta}\right) - E\left(W\left(\hat{\theta}\right)\right)}{\sqrt{Var\left(W\left(\hat{\theta}\right)\right)}} \sim N\left(0,1\right).$$

Null distribution of $l_{zm}^{(\mathrm{C})}$ and $l_{zm}^{*(\mathrm{C})}$

Conijn et al.'s (2014) approach is similar to Drasgow's et al. (1991) approach, but the order of operations is reversed. That is, $l_{z(s)}$ is computed for each subtest *s* and then

all values are added. $l_{z(s)}$ is asymptotically standard normally distributed for true θ values. Furthermore, due to the local independence assumption, the $l_{z(s)}$ statistics are independent. As a consequence, $l_{zm}^{(C)}$ is the sum of *S* independent standard normally distributed variables and is therefore normally distributed with mean and variance equal to the sums of the means and variances, respectively. Thus, the asymptotic null distribution of $l_{zm}^{(C)}$, assuming known θ and local independence, is given by

$$l_{zm}^{(C)} = \sum_{s=1}^{S} l_{z(s)} \sim N(0, S).$$

The asymptotic distribution of $l_{zm}^{*(C)}$ can be derived along the same lines. Since $l_{zm}^{*(C)}$ is the sum of the $l_{z(s)}^{*}$ values, each independent and asymptotically *N*(0, 1) distributed (Snijders, 2001), we immediately have that

$$l_{zm}^{*(C)} \sim N(0, S)$$

for trait estimates $\hat{\theta}_{(s)}$ satisfying Equation 7.

Design of the simulation study

In practice, to assess whether response patterns are unusual one can either (1) simulate a large number of response patterns under the null distribution of normal behavior and compare the observed with the simulated response patterns; or (2) compute the critical value on the basis of the asymptotic distribution. Obviously, the first approach is time-consuming but has the benefit of not having to rely on asymptotic theory. The second approach is computationally much more efficient but does rely on asymptotic theory.

The main goal of this simulation study was to study the quality of the asymptotic results discussed in the previous sections. In particular, we wanted to verify how the asymptotic approximations for person-fit statistics $l_{zm}^{*(\mathrm{D})}$ and $l_{zm}^{*(\mathrm{C})}$ hold for relatively short subtest lengths (say, of 10 items). The goal is to understand whether the asymptotic results are accurate enough for most practical purposes. Discussing the univariate l_z^* statistic, Snijders (2001, p. 332) expected that $n \geq 15$ would be sufficient for the asymptotic approximations to work well (in case of univariate scales). Subtests might be of shorter length than the 15 items mentioned by Snijders. How much shorter the subtests can be is dependent on the relation between the subtests. If the latent traits for the subtests correlate perfectly (i.e., test taker's θ is the same for each subtest), the test is actually univariate and, according to Snijders, subtests of lengths $n_i \approx 15/S$ should suffice. When the correlation between subtest traits is smaller than one, one may expect to need subtests of longer length. Studying how subtest length and trait correlations relate to the quality of the asymptotic approximations is the main goal of this simulation study.

The simulation study was set up as follows. Item scores of 1,000 test takers on four subtests were generated. Four subtest lengths were considered: 10, 25, 50, and 100. The shorter subtest lengths (10, 25) are of most practical interest, whereas the longer subtest lengths (50, 100) are mostly of theoretical interest. All subtests within the same dataset had the same length. The 2PLM was used to generate the item scores, with discrimination parameters uniformly distributed between [0.5, 2.0] and difficulty parameters standard normally distributed (bounded between –2.5 and +2.5). Moreover, four person θ parameters were generated for each simulated test taker, one per subtest. These parameters were randomly drawn from a multivariate normal distribution. Seven

between-subtests correlations of θ were considered: 0.4(0.1)1.0. These item and person parameters resulted in data that were very similar to the empirical data from a number of large scale high-stakes educational admission tests (see also Rupp, 2013).

The simulation study consisted therefore of a 4 (number of subtest lengths) by 7 (number of between-subtests correlations of θ) completely crossed design, hence 28 experiment conditions in total. One hundred replications were simulated per condition. For each replicated dataset, six multiple subtests person-fit statistics were computed. Of these, $l_{zm}^{*(\mathrm{D})}$ and $l_{zm}^{*(\mathrm{C})}$, which we proposed and developed in this report, were of most interest. Furthermore, we also computed $l_{zm}^{(\mathrm{D})}$ and $l_{zm}^{(\mathrm{C})}$ in order to compare these uncorrected statistics with their starred versions. We expected the corrected starred statistics to outperform the uncorrected statistics. Finally, we computed l_z and l_z^* by concatenating the four subtests together, that is, by ignoring the multiple subtests data structure. We expected this approach to work well for large correlation values between the θs but not so well for lower correlation values between the θs.

The simulation was coded in R (R Core Team, 2014). The item parameters were estimated by means of the function est() in the 'irtoys' package (Partchev, 2014). The maximum likelihood person parameters were estimated by means of the function mlebme(), also in the 'irtoys' package.

Results of the simulation study

Findings are reported in various tables and figures. For l_z^*, $l_{zm}^{*(\mathrm{C})}$, and $l_{zm}^{*(\mathrm{D})}$, Table 1 lists the following values: (i) the mean of the 1,000 statistic values per replication, averaged across the 100 replications; (ii) the standard deviation of the 1,000 statistic values per replication, averaged across the 100 replications; (iii) the Kolmogorov-

Smirnov (KS) distance between the empirical and theoretical (asymptotic) normal cumulative distribution function; and (iv) the level of significance when applying critical values from the asymptotic distribution, at $\alpha = 0.05$. The Kolmogorov-Smirnov distance (Smirnov, 1948) is a method to assess whether the empirical results lie close to the asymptotic distribution. This metric is a common method for density comparisons and reports the maximum vertical distance between both cumulative distributions. When both distributions completely agree, this value is zero; when they completely disagree, it is one. For l_z, $l_{zm}^{(C)}$, and $l_{zm}^{(D)}$, Table 2 lists the means and standard deviations over the replications. (Reporting KS-distances and levels of significance for these statistics is undesirable, since the asymptotic distribution only holds if all θ are known.)

We first focus on the results in Table 1. The values for the means and standard deviations can be directly compared with the means and standard deviations of the asymptotic distribution. With respect to the means, Table 1 shows that (i) the empirical means are structurally larger than zero across all methods; and that (ii) the means decrease as n_i increases. Furthermore, the value of the means seem unrelated to the value of the subtest correlations ρ, with the exception of l_z^* which seems to have slightly larger mean values for larger ρ (for instance, for $n_i = 100$, the means range from 0.047 ($\rho = 0.4$) through 0.076 ($\rho = 1$)). With respect to the standard deviations, Table 1 shows that the empirical sd's are very close to their asymptotic values (1 for l_z^* and $l_{zm}^{*(D)}$; 2 for $l_{zm}^{*(C)}$) across all methods.

Table 1 shows that l_z^* is sensitive to ρ: The Kolmogorov-Smirnov-distance increases when ρ decreases, especially for larger subtest lengths. This result is to be expected, as the idea of ignoring the multiple subtests structure is incompatible with

increasingly lower values of correlations between the θs. The two multiple subtest methods, $l_{zm}^{*(C)}$ and $l_{zm}^{*(D)}$, do not show this dependence on ρ. For all methods it holds that, when n_i increases, the asymptotic approximation lies closer to the empirical distribution. Furthermore, the Kolmogorov-Smirnov distances decrease when the subtest length increases. This is obvious: faced with more data, better predictions can be made.

The Kolmogorov-Smirnov-distance measures how close the complete empirical distribution is with respect to the asymptotic distribution. This is actually more than what we need: What happens in the critical region of the distribution (i.e., the lower-tail) is what is important when looking for aberrant patterns. We are not (primarily) interested in whether the l_z^*-scores of fitting response patterns is estimated without bias; what matters most is that the scores for misfitting response patterns are measured accurately. Figure 1, based on the $l_{zm}^{*(D)}$ values for the experiment condition defined by $n_i = 25$ and $\rho = 0.7$, displays the empirical and asymptotic density functions (left) and cumulative distribution functions (right), with the 1% and 5% critical values added. The full empirical distribution has a significant misfit compared with the asymptotic standard normal because of its skew to the right. The skewness of this distribution has been noted in practice as well (e.g., Meijer & Tendeiro, 2012, their Figure 1). However, the left tail of the empirical distribution is relatively well approximated by the asymptotic distribution, especially at the 1% level. In Table 1 we report the α_{asymp} values, which consist of the proportion of empirical data to the left of the 5% quantile of the asymptotic distribution. Thus, α_{asymp} describes for what proportion of the empirical results the null hypothesis of no aberrant behavior would be rejected (a Type I error), if this decision is made based on asymptotic theory and $\alpha = 0.05$. Values close to 0.05 are indicative of the adequacy of the

asymptotic approximation. Figure 2 presents a visualization of the same results. For comparison, Figure 2 also shows the α_{asymp} values for the statistics without Snijders' bias-correction, where the asymptotic distribution is derived under the additional (and incorrect) assumption of known θ.

From Figure 2, we can conclude the following. (i) One should not rely on asymptotic theory for the non-bias removed statistics (right panels): Even for large subtests (n_i = 50) the α_{asymp} values can be less than half the nominal values (around 2%). Thus, the critical values based on the asymptotic approximation are too conservative in the case of non-corrected statistics (i.e., there is lack of power). The problem is also present for the bias-corrected statistics (left panels) but to a lesser extent. (ii) l_z^* works very well for very small subtests (n_i = 10), but for lower correlations and lengthier subtests the critical values from the asymptotic distribution yield are too liberal: Too many response patterns are flagged as aberrant. (iii) The performance of $l_{zm}^{*(C)}$ and $l_{zm}^{*(D)}$ is comparable and both methods are unaffected by the value of ρ. (iv) Even for subtests of moderate length (n_i = 25), the approximation by asymptotic theory provides accurate approximation.

Deviations between the reported α_{asymp} values and the nominal α = 0.05 can be due to (a combination of) two reasons: (i) Sampling variation (results are based on 100 replications of 1,000 simulated persons), and (ii) the approximation is asymptotic and the sample size is clearly finite. However, sampling variation was controlled almost entirely by our experiment design. When sampling 100 × 1,000 = 100,000 values from a normal distribution, then in 95% of cases, the α_{asymp} value would be in (0.0499 , 0.05001).

Table 3 uses formal regression models to underpin the conclusions of Tables 1 and 2 and Figure 2. For each of the six types of person-fit-statistic, first the regression model $\text{mean}_j = \beta_0 + \beta_1(n_i)_j + \beta_2\rho_j + \beta_3(n_i \times \rho_j) + \varepsilon_j$ is fitted to the 4 × 7 combinations of subtest length n_i and subtest correlation ρ and the p-values and effect sizes are reported. Next, a similar model, now with α_{asymp} as the dependent variable, is fitted for the three starred methods. We decided to include an interaction term since Tables 1 and 2 and Figure 2 indicate that such interaction might be present. It has to be noted that this elementary linear model is not perfect; especially the subtest lengths seem to have a non-linear relation with the dependent variable. However, the model seems adequate for a rough indication. The results from the table are clear: for every method, the size of the subtest is a significant factor with large effect sizes. The subtest correlation is only significant and relevant for l_z^* and l_z: the multiple subtests approaches indeed are capable of dealing with correlated subtests without distortions in the mean l_z^* or l_z value nor their corresponding α_{asymp} value. For the α_{asymp} values, a clear interaction is present only for l_z^*, for the mean values, none of the interactions are significant (at the usual 5% level) nor do they have considerable effect sizes.

Discussion

In this paper we investigated the theoretical asymptotic distributions of three person-fit statistics for tests that consists of multiple subtests. In both psychological and educational measurement these types of tests are often used, but thus far there were no studies that investigated these asymptotic distributions. A recent study that used the multiple subtest extensions $l_{zm}^{*(C)}$ and $l_{zm}^{*(D)}$ made use of simulation to determine the critical values on the basis of which item score patterns could be classified as normal or aberrant

(Tendeiro et al., 2014). A drawback of this approach is that it is time-consuming. In the present study we showed that asymptotic theory can adequately be used for both statistics even for subtest lengths as low as 10 items. and that, at least, at for a 95% confidence interval type I errors are in agreement with what is expected.

Type I errors are controlled for by the simple univariate l_z^* statistic when correlations between tests are relatively high (larger than, say .7-.8). This is the case for many large-scale educational tests. Drasgow et al. (1991), for example, reported a correlation r = .73 between SAT Verbal and Quantitative tests and r = .80 between the enhanced ACT Engish and Mathematics test. Thus, in these cases the theoretical asymptotic distribution of both the l_z^* statistic and the multiple subtests extensions $l_{zm}^{*(\mathrm{C})}$ and $l_{zm}^{*(\mathrm{D})}$ can be used. As many studies showed (e.g., Conijn et al., 2014), correlations between test scores for non-cognitive instruments are often lower than for cognitive tests. In these cases the asymptotic distribution of both $l_{zm}^{*(\mathrm{C})}$ or $l_{zm}^{*(\mathrm{D})}$ can be used, at least for α = 0.05. We should note, however, that as we showed in Figure 1, because the empirical distributions are skewed, at an α level of, for example, α = 0.10, results may be less optimal. On the other hand, almost all person-fit statistics use α levels of .05 or lower, so in practice we conclude that this study showed that researchers can use the discussed asymptotic distributions to classify item score patterns as normal or aberrant for multiple subtests settings.

There can be benefit in applying bootstrap methods (such as those in Tendeiro et al., 2014) rather than resorting to asymptotic theory. These benefits especially hold in small studies, when the use of asymptotic theory is questionable. However, in our paper we show that, even for fairly short subtest lengths, asymptotic results already provide

decent approximations. Finally, the benefit of not having to use the bootstrap distribution is saving computing time and, to a lesser degree, it is less technical: understanding the bootstrap is quite hard; flagging all l_z^* scores below -1.65 is extremely simple.

Supplementary materials

The *R* code used to generate the results, figures and tables, as well as a detailed version of Figure 2, is provided as online supplementary material.

References

Armstrong, R. D., Stoumbos, Z. G., Kung, M. T., & Shi, M. (2007). On the performance of the l_z person-fit statistic. *Practical Assessment, Research & Evaluation, 12 (16).* Retrieved from http://pareonline.net/getvn.asp?v=12&n=16

Conijn, J. M., Emons, W. M., & Sijtsma, K. (2014). Statistic lz-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement,* 38, 122-136. doi:10.1177/0146621613497568

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38,* 67-86. doi:10.1111/j.2044-8317.1985.tb00817.x

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15,* 171-191. doi:10.1177/014662169101500207

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's l_z^* index of person fit with emphasis on response model selection and ability estimation.

Journal of Educational and Behavioral Statistics, 37, 57-81, doi: 10.3102/1076998610396894

Partchev, I. (2014). *irtoys: Simple interface to the estimation and plotting of IRT models*. R package version 0.1.7. http://CRAN.R-project.org/package=irtoys

Smirnov, N. (1948) Table for estimating the goodness of fit of empirical distributions. *Annals of mathematical statistics*, *19*, 279-281. doi: 10.1214/aoms/1177730256

Snijders, T. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*, 331-342. doi:10.1007/BF02294437

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135. doi:10.1177/01466210122031957

Meijer, R.R., & Tendeiro, J. N. (2012). The use of lz and lz* person-fit statistics and problems derived from model misspecification. *Journal of Educational and Behavioral Statistics, 37*, 758-766.

Meijer, R. R., & Tendeiro, J. N. (2014). *The use of person-fit scores in high-stakes educational testing: How to use them and what they tell us* (LSAC Research Report, 14-03). Newtown, PA: Law School Admission Council.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessment Modeling*, *55*, 3-38.

Tendeiro, J. N., Meijer, R. R., & Albers, C. J. (2014). *Detection of invalid test scores on admission tests: A simulation study using person-fit statistics* (LSAC Research Report, submitted). Newtown, PA: Law School Admission Council.

van Krimpen-Stoop, E. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 327-345. doi:10.1177/01466219922031446

Table 1

Results of the simulation study for person-fit statistics l_z^*, $l_{zm}^{*(C)}$, and $l_{zm}^{*(D)}$: Mean, standard deviation, Kolmogorov-Smirnov (KS) distances, and empirical proportion of statistic values scoring below the 5% quantile of the asymptotic distribution). Results are averaged across replications.

		l_z^*				$l_{zm}^{*(C)}$				$l_{zm}^{*(D)}$			
ρ	n_i	mean	sd	KS	α_{asymp}	mean	sd	KS	α_{asymp}	mean	sd	KS	α_{asymp}
0.4	10	0.083	1.017	0.076	0.056	0.642	1.999	0.166	0.037	0.336	0.996	0.176	0.036
	25	0.053	1.067	0.059	0.064	0.415	2.006	0.110	0.040	0.212	1.002	0.113	0.040
	50	0.045	1.142	0.064	0.077	0.312	2.005	0.086	0.042	0.155	1.001	0.087	0.042
	100	0.047	1.269	0.084	0.096	0.264	2.008	0.072	0.042	0.129	1.002	0.072	0.042
0.5	10	0.079	1.009	0.071	0.055	0.609	2.016	0.157	0.038	0.318	1.003	0.167	0.037
	25	0.055	1.054	0.057	0.062	0.413	2.012	0.110	0.040	0.211	1.005	0.113	0.040
	50	0.048	1.108	0.059	0.070	0.313	1.999	0.087	0.042	0.155	0.998	0.087	0.042
	100	0.052	1.220	0.074	0.085	0.265	2.003	0.073	0.042	0.130	1.000	0.072	0.041
0.6	10	0.082	1.008	0.072	0.055	0.622	2.009	0.160	0.037	0.322	1.002	0.169	0.038
	25	0.056	1.038	0.056	0.058	0.413	2.009	0.111	0.040	0.211	1.004	0.114	0.041
	50	0.051	1.075	0.053	0.063	0.314	2.002	0.086	0.041	0.156	0.999	0.086	0.042
	100	0.056	1.165	0.065	0.076	0.263	1.997	0.071	0.042	0.130	0.997	0.071	0.041
0.7	10	0.083	1.003	0.072	0.054	0.620	2.002	0.160	0.037	0.326	0.998	0.172	0.037
	25	0.058	1.020	0.053	0.055	0.411	2.004	0.109	0.040	0.210	1.000	0.112	0.040
	50	0.054	1.048	0.050	0.059	0.315	2.001	0.085	0.041	0.156	0.999	0.085	0.041
	100	0.061	1.115	0.055	0.065	0.262	2.005	0.071	0.042	0.130	1.000	0.071	0.041
0.8	10	0.084	0.999	0.073	0.053	0.625	2.010	0.161	0.037	0.326	1.000	0.172	0.037
	25	0.060	1.010	0.054	0.054	0.413	2.008	0.110	0.041	0.211	1.003	0.112	0.040
	50	0.056	1.030	0.050	0.055	0.313	2.014	0.087	0.043	0.156	1.006	0.087	0.043
	100	0.065	1.074	0.049	0.058	0.262	2.008	0.071	0.042	0.130	1.003	0.071	0.042
0.9	10	0.084	0.997	0.073	0.053	0.615	2.005	0.160	0.038	0.320	0.999	0.168	0.037
	25	0.062	0.999	0.055	0.052	0.417	2.000	0.111	0.040	0.212	1.000	0.113	0.040
	50	0.060	1.008	0.048	0.050	0.315	2.007	0.086	0.042	0.157	1.002	0.086	0.042
	100	0.071	1.048	0.046	0.051	0.263	2.008	0.073	0.042	0.130	1.003	0.073	0.043
1.0	10	0.084	1.001	0.075	0.054	0.614	2.015	0.161	0.038	0.319	1.005	0.169	0.038
	25	0.062	1.002	0.054	0.053	0.410	2.011	0.110	0.041	0.208	1.006	0.111	0.042
	50	0.063	1.001	0.048	0.049	0.316	2.000	0.085	0.041	0.158	0.999	0.085	0.041
	100	0.076	1.039	0.047	0.048	0.263	2.007	0.072	0.042	0.131	1.003	0.072	0.042

Note: ρ = Correlation between subtest θ values; n_i = Subtest length.

Table 2

Results of the simulation study for person-fit statistics l_z, $l_{zm}^{(C)}$, and $l_{zm}^{(D)}$: Mean and standard deviation. Results are averaged across replications.

		l_z		$l_{zm}^{(C)}$		$l_{zm}^{(D)}$	
ρ	n_i	mean	sd	mean	sd	mean	sd
0.4	10	0.070	0.912	0.543	1.595	0.267	0.780
	25	0.039	0.956	0.353	1.673	0.173	0.818
	50	0.029	1.028	0.260	1.699	0.129	0.832
	100	0.025	1.146	0.217	1.719	0.108	0.844
0.5	10	0.065	0.885	0.508	1.582	0.249	0.770
	25	0.040	0.937	0.351	1.679	0.173	0.822
	50	0.032	0.990	0.261	1.697	0.129	0.833
	100	0.029	1.093	0.218	1.720	0.108	0.846
0.6	10	0.069	0.882	0.519	1.584	0.255	0.773
	25	0.043	0.914	0.352	1.674	0.173	0.821
	50	0.036	0.952	0.263	1.702	0.130	0.837
	100	0.033	1.030	0.217	1.712	0.108	0.843
0.7	10	0.069	0.874	0.524	1.582	0.258	0.773
	25	0.044	0.890	0.349	1.667	0.172	0.818
	50	0.038	0.918	0.261	1.696	0.130	0.835
	100	0.038	0.972	0.217	1.713	0.108	0.846
0.8	10	0.071	0.866	0.526	1.587	0.259	0.776
	25	0.048	0.878	0.352	1.675	0.174	0.825
	50	0.042	0.899	0.263	1.712	0.130	0.846
	100	0.043	0.927	0.217	1.723	0.108	0.853
0.9	10	0.072	0.856	0.516	1.583	0.254	0.775
	25	0.052	0.864	0.356	1.672	0.177	0.825
	50	0.046	0.873	0.263	1.706	0.131	0.846
	100	0.048	0.890	0.216	1.722	0.108	0.855
1.0	10	0.074	0.851	0.517	1.587	0.255	0.779
	25	0.053	0.857	0.345	1.670	0.172	0.827
	50	0.050	0.861	0.264	1.700	0.131	0.846
	100	0.054	0.874	0.216	1.719	0.108	0.857

Note: ρ = Correlation between subtest θ values; n_i = Subtest length.

Table 3

Effect sizes (η^2) and p-values for the regression models predicting the mean values and α_{asymp} values, with sample size and subtest correlation as predictors.

		l_z^*		$l_{zm}^{*(C)}$		$l_{zm}^{*(D)}$		l_z		$l_{zm}^{(C)}$		$l_{zm}^{(D)}$	
		p	η^2	*p*	η^2	*p*	η^2	*p*	η^2	*p*	η^2	*p*	η^2
mean	n_i	.013	.182	.000	.726	.000	.720	.000	.417	.000	.741	.000	.742
	ρ	.021	.152	.933	.000	.939	.000	.003	.171	.933	.000	.985	.000
	$n_i \times \rho$	.121	.065	.938	.000	.839	.000	.145	.036	.927	.000	.947	.000
α_{asymp}	n_i	.000	.227	.000	.584	.000	.540						
	ρ	.000	.476	.792	.001	.593	.006						
	$n_i \times \rho$	.000	.270	.962	.000	.792	.001						

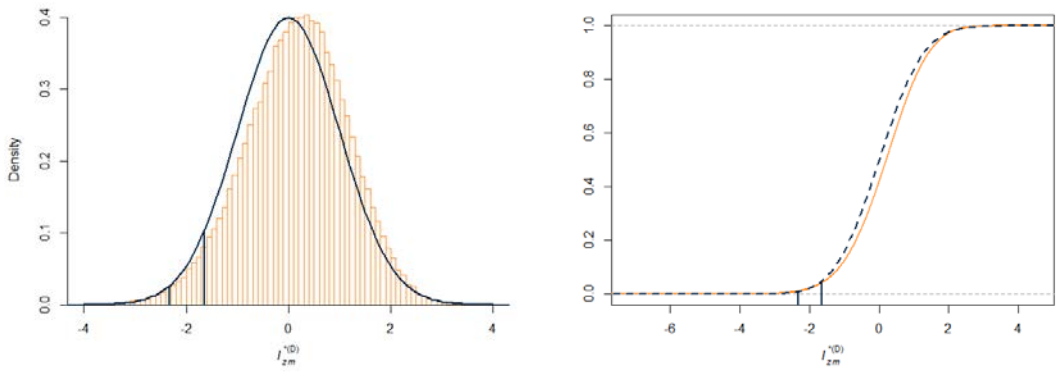

Figure 1. Left: Histogram of the 1,000 × 100 = 100,000 computed $l_{zm}^{*(D)}$ values (orange) and the standard normal distribution (blue). The blue vertical lines correspond to the α = 1% (left) and α = 5% (right) critical values. Right: The corresponding cumulative distribution functions (solid orange:empirical; dashed blue: standard normal). It can be seen that the maximal vertical distance in the r.h.s. display occurs around $l_{zm}^{*(D)} \approx 0$. This figure is based on the experiment condition defined by the parameters n_i = 25 and ρ = 0.7.

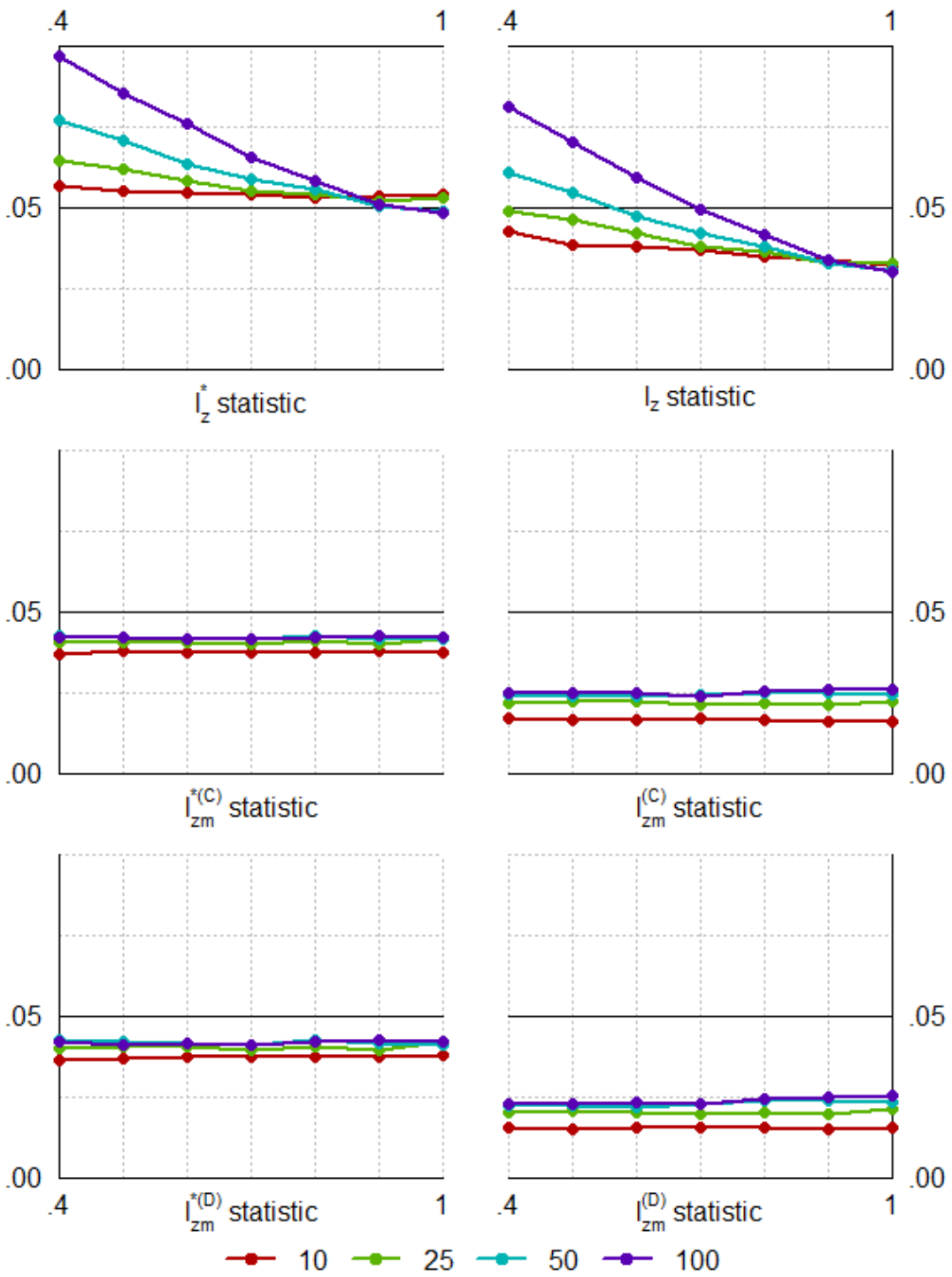

Figure 2. The α_{asymp} values of the 100 replications of the 1,000 person-fit statistics. Left panels, from top to bottom: l_z^*, $l_{zm}^{*(C)}$, and $l_{zm}^{*(D)}$. Right panels, from top to bottom: l_z, $l_{zm}^{(C)}$, and $l_{zm}^{(D)}$. In the online supplementary material, a more detailed version of this Figure is provided.