This is a repository copy of *Critical Values for Yen's Q3: Identification of Local Dependence in the Rasch model using Residual Correlations*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/106017/

Version: Accepted Version

Critical Values for Yen's Q_3: Identification of Local Dependence in the Rasch model using Residual Correlations.

Karl Bang Christensen

Section of Biostatistics, Department of Public Health, University of Copenhagen

Copenhagen

Denmark

E-mail: KACH@sund.ku.dk

Guido Makransky, PhD

Department of Psychology, University of Southern Denmark

Odense

Denmark

E-mail: gmakransky@health.sdu.dk

Mike Horton

Academic Department of Rehabilitation Medicine, Leeds Institute of Rheumatic and Musculoskeletal Medicine, University of Leeds

Leeds

UK

Email: M.C.Horton@leeds.ac.uk

Abstract

The assumption of local independence is central to all IRT models. Violations can lead to inflated estimates of reliability and problems with construct validity. For the most widely used fit statistic Q_3 there are currently no well-documented suggestions of the critical values which should be used to indicate local dependence, and for this reason a variety of arbitrary rules of thumb are used. In this study, we used an empirical data example and Monte Carlo simulation to investigate the different factors that can influence the null distribution of residual correlations, with the objective of proposing guidelines that researchers and practitioners can follow when making decisions about local dependence during scale development and validation. We propose that a parametric bootstrapping procedure should be implemented in each separate situation in order to obtain the critical value of local dependence applicable to the data set, and provide example critical values for a number of data structure situations. The results show that for the Q_3 fit statistic no single critical value is appropriate for all situations, as the percentiles in the empirical null distribution are influenced by the number of items, the sample size, and the number of response categories. Furthermore, our results show that local dependence should be considered relative to the average observed residual correlation, rather than to a uniform value, as this results in more stable percentiles for the null distribution of an adjusted fit statistic.

Keywords: Local dependence, Rasch model, Yen's Q_3, Residual correlations, Monte Carlo simulation.

Introduction

Statistical independence of two variables implies that knowledge about one variable does not change our expectations about another variable. Thus, test items $X_1,…, X_I$ are not independent, since a student's giving a correct answer to one test item would change our expectation of her probability of also giving a correct answer to another item in the same test. A fundamental assumption in the Rasch (1960) model and in other IRT models is that item responses are conditionally independent given the latent variable

$$P(X_1 = x_1, …, X_I = x_I | \theta) = \prod_{i=1}^{I} P(X_i = x_i | \theta). \qquad (1)$$

The items should only be correlated through the latent trait that the test is measuring (Lord and Novick, 1968). This is generally referred to as local independence (Lazarsfeld & Henry, 1968).

The assumptions of local independence can be violated through response dependency and multidimensionality, and these violations are often referred to under the umbrella-term of 'local dependence' (LD). Both of these situations yield inter-item correlations beyond what can be attributed to the latent variable, but for very different reasons. Response dependency occurs when items are linked in some way, such that the response on one item governs the response on another because of similarities in, for example, item content or response format. A typical example is where several walking items are included in the same scale. If a person can walk several miles without difficulty, then that person must be able to walk one mile, or any lesser distance, without difficulty (Tennant and Conaghan, 2007). This is a structural dependency which is inherent within the items, because there is no other logical way in which a person may validly respond. Another form of LD could be caused by a redundancy-dependency, where the degree of overlap within the content of items is such that the items are not independent (i.e. where the same question is essentially asked

twice, using slightly different language or synonymous descriptive words). Yen (1993) offers an in depth discussion of ways that the format and presentation of items can cause LD.

Violation of the local independence assumption through multidimensionality is typically seen for instruments composed of bundles of items that measure different aspects of the latent variable, or different domains of a broader latent construct. In this case the higher order latent variable alone might not account for correlations between items in the same bundle.

Violations of local independence in a unidimensional scale will influence estimation of person parameters and can lead to inflated estimates of reliability and problems with construct validity. Consequences of LD have been described in detail elsewhere (Yen 1993; Scott and Ip, 2002; Lucke, 2005; Marais, and Andrich, 2008b; Marais, 2009). Ignoring LD in a unidimensional scale thus leads to reporting of inflated reliability giving a false impression of the accuracy and precision of estimates (Marais, 2013). For a discussion of the effect of multidimensionality on estimates of reliability see Marais and Andrich (2008a).

Detecting Local Dependence

One of the earliest methods for detecting local dependence in the Rasch model is the fit measure Q_2 (van den Wollenberg, 1982), which was derived from contingency tables and used the sufficiency properties of the Rasch model. Kelderman (1984) expressed the Rasch model as a log linear model in which LD can be shown to correspond to interactions between items. Loglinear Rasch models have also been considered by Haberman (2007) and by Kreiner and Christensen (2004, 2007), who proposed to test for LD by evaluating partial correlations using approach similar to the Mantel-Haenszel analysis of DIF (Holland and Thayer, 1988). The latter approach is readily implemented in standard software like SAS or SPSS. Notably, Kreiner and Christensen (2007) argue that the log linear Rasch models proposed by Kelderman (1984) that incorporate LD still provide essentially

valid and objective measurement and describe the measurement properties of such models.

Furthermore, a way of quantifying local dependence has been proposed by Andrich and Kreiner (2010) for two dichotomous items. It is based on splitting a dependent item into two new ones, according to the responses to the other item within the dependent pair. Local dependence is then easily quantified by estimating the difference d between the item locations of the two new items. However, Andrich and Kreiner do not go on to investigate if d is statistically significant. For the partial credit model (Masters, 1982) and the rating scale model (Andrich, 1978) a generalized version this methodology exists (Andrich, Humphry and Marais, 2012)

Beyond the Rasch model, Yen (1984) proposed the Q_3 statistic for detecting LD in the 3PL model. This fit statistic is based on the item residuals

$$d_i = X_i - E(X_i|\hat{\theta}). \qquad (2)$$

and computed as the Pearson correlation (taken over examinees)

$$Q_{3,ij} = r_{d_i d_j} \qquad (3)$$

where d_i and d_j are item residuals for items i and j, respectively. This method is often used for the Rasch model, the partial credit model and the rating scale model.

Chen and Thissen (1997) discussed X^2 and G^2 LD statistics that, while not more powerful than the Q_3, have null distributions very similar to the chi-squared distribution with one degree of freedom. Other methods for detecting LD are standardized bivariate residuals for dichotomous (Reiser, 1996) or multinomial IRT models (Maydeu-Olivares and Liu, 2015), the use of conditional covariances (Douglas et al, 1998), or the use of Mantel-Haenzsel type tests (Ip, 2001). Tests based on parametric models are also a possibility: Glas and Suarez-Falcon (2003) proposed Lagrange multiplier (LM) tests based on a threshold shift model, but bifactor models (Liu and Thissen, 2012; 2014), specification of other models that incorporate local dependence (Hoskens and De Boeck,

1997; Ip, 2002), or limited information goodness-of-fit tests (Liu and Maydeu-Olivares, 2013) is also possible.

The use of the Q_3 fit statistic

Yen's Q_3 is probably the most often reported index in published Rasch analyses due to its inclusion (in the form of the residual correlation matrix) in widely used software like RUMM (Andrich, Sheridan and Luo, 2010). Yen (1984) argued that if the IRT model is correct then the distribution of the Q_3 is known, and proposed that p-values could be based on the Fisher (1915) z-transform. Chen and Thissen (1997) stated: "In using Q_3 to screen items for local dependence, it is more common to use a uniform critical value of an absolute value of 0.2 for the Q_3 statistic itself". They went on to present results showing that, while the sampling distribution under the Rasch model is bell shaped, it is not well approximated by the standard normal, especially in the tails (Chen and Thissen, 1997, Figure 3).

In practical applications of the Q_3 test statistic researchers will often compute the complete correlation matrix of residuals and look at the maximum value

$$Q_{3,max} = \max_{i \neq j} Q_{3,ij}. \qquad (4)$$

Critical Values of Residual Correlations

When investigating LD based on Yen's Q_3, residuals for any pair of items should be uncorrelated, and generally close to 0. Residual correlations that are high indicate a violation of the local independence assumption, and this suggests that the pair of items have something more in common than the rest of the item set have in common with each other (Marais, 2013).

As noted by Yen (1984, p.127) a negative bias is built into Q3. This problem is due to the fact that measures of association will be biased away from zero even though the assumption of local

independence applies, due to the conditioning on a proxy variable instead of the latent variable (Rosenbaum, 1984). A second problem is that the way the residuals are computed induce a bias (Kreiner and Christensen, 2011). Marais (2013) recognized that the sampling properties among residuals are unknown; therefore these statistics cannot be used for formal tests of LD. A third, and perhaps the most important, problem in applications, is that there are currently no well-documented suggestions of the critical values which should be used to indicate LD, and for this reason arbitrary rules of thumb are used when evaluating whether an observed correlation is such that it can be reasonably supposed to have arisen from random sampling.

Standards often reported in the literature include looking at fit residuals over the critical value of 0.2, as proposed by Chen and Thissen (1997). For examples of this see, e.g., Reeve et al. 2007; Hissbach, Klusmann and Hampe, 2011; Makransky and Bilenberg, 2014; Makransky, Rogers and Creed, 2014. However, other critical values are also used, and there seems to be a wide variation in what is seen as indicative of dependence. Marais and Andrich (2008b) investigated dependence at a critical residual correlation value of 0.1, but a value of 0.3 has also often been used (see e.g. La Porta et al., 2011; Das Nair et al., 2011; Ramp et al. 2009; Røe, et al. 2014), and critical values of 0.5 (ten Klooster et al. 2008; Davidson et al. 2004) and even 0.7 (González-de Paz et al., 2014) can be found in use.

There are two fundamental problems with this use of standard critical values: (i) there is limited evidence of their validity and often no reference of where values come from, and (ii) they are not sensitive to specific characteristics of the data.

Marais (2013) identified that the residual correlations are difficult to directly interpret confidently when there are fewer than 20 items in the item set, but also stated that the correlations should always be considered relative to the overall set of correlations. This is because the magnitude of a residual correlation value which indicates LD will vary depending on the number of

items in a data set. Instead of an absolute critical value, Marais (2013) suggests that residual

correlation values should be compared to the average item residual correlation of the complete data

set to give a truer picture of the LD within a data set. It was concluded that when diagnosing

response dependence, item residual correlations should be considered relative to each other and in

light of the number of items, although there is no indication of a relative critical value (above the

average residual correlation) that could indicate LD.

Thus, under the null hypothesis the average correlation of residuals is negative (cf. Marias

(2013, p.121)) and, ideally, observed correlations between residuals in a data set should be

evaluated with reference to this average value. Marais proposes to evaluate them with reference to

the average of the observed correlations rather than the average under the null hypothesis. Thus,

following Marais, we could consider the average value of the observed correlations

$$\bar{Q}_3 = \binom{I}{2}^{-1} \sum_{i \neq j} Q_{3,ij} \qquad (5)$$

where $\binom{I}{2}$ is the number of item pairs and define the test statistic

$$Q_{3,*} = Q_{3,max} - \bar{Q}_3 \qquad (6)$$

that compares the largest observed correlation to average of the observed correlations.

The problem with the currently used critical values is that they are neither theoretically nor

empirically based. Researchers and practitioners faced with making scale validation and

development decisions need to know what level of LD could be expected, given the properties of

their items and data.

A possible solution would be to use a parametric bootstrap approach and simulate the residual

correlation matrix several times under the assumption of fit to the Rasch model. This would provide

information about the level of residual correlation that could be expected for the particular case,

given that the Rasch model fits. To our knowledge, there is no existing research that describes how

important characteristics such as the number of items, number of response categories, number of

respondents, the distribution of items and persons, and the targeting of the items impact residual correlations expected, given fit to the Rasch model. In the current study we investigate the possibility of identifying critical values of LD by examining the distribution of Q_3 under the null hypothesis, where the data fits the model. This is done using an empirical example along with a simulation study.

Given the existence of the wide range of fit statistics with known sampling distributions outlined above it is surprising that Rasch model applications abound with reporting of Q3 using arbitrary cut-points without theoretical or empirical justification. The reason for this is that the theoretically sound LD indices are not included in the software packages used by practitioners. For this reason this article presents extensive simulation studies that will (a) illustrate that Q3 should be interpreted with caution (b) allow researchers to know what level of LD could be expected, given properties of their items and data. Furthermore these simulation studies will be used to study if the maximum correlation or the difference between the maximum correlation and the average correlation as suggested by Marais (2013). Thus, the objectives of this paper are: (i) to provide an overview of the influence of different factors upon the null distribution of residual correlations, (ii) to propose guidelines that researchers and practitioners can follow when making decisions about LD during scale development and validation. Two different situations are addressed: firstly, the situation where the test statistic is computed for all item pairs and only the strongest evidence (the largest correlation) considered, and secondly, the less common case, where only a single a priori defined item pair is considered.

Simulation study

Methods: Simulation Study

The simulated data sets used (i) I dichotomous items simulated from

$$P(X_i = x|\theta) = \frac{\exp(x(\theta - \beta_i))}{1 + \exp(\theta - \beta_i)} \qquad (i = 1, \dots, I) \qquad (7)$$

with evenly spaced item difficulties β_i ranging from -2 to 2

$$\beta_i = 2\left(\frac{i-1}{I-1}\right) \qquad (i = 1, \dots, I) \qquad (8)$$

or (ii) I polytomous items simulated from

$$P(X_i = x|\theta) = \frac{\exp(\sum_{h=1}^{x}(\theta - \beta_{ih}))}{1 + \sum_{l=1}^{3}\exp(\sum_{h=1}^{l}(\theta - \beta_{ih}))} \qquad (x = 0,1,2,3;\ i = 1, \dots, I) \qquad (9)$$

with item parameters defined by

$$\beta_{ih} = 2\left(\frac{i-1}{I-1}\right) + (h - 1) \qquad (i = 1, \dots, I;\ h = 1,2,3) \qquad (10)$$

The person locations were simulated from a normal distribution with mean μ and SD 1. All combinations of the four conditions: (a) number of items (I = 10, 15, 20); (b) number of persons (N =200, 250, … , 1000); (c) number of response categories (two, four); and (d) mean value in the distribution of the latent variable θ (μ = 0, 2) were simulated. This yielded 204 different setups, and for each of these we simulated 10,000 data sets and followed the steps

(i) estimating item parameters using pairwise conditional estimation (Zwinderman, 1995; Andrich and Luo, 2003),

(ii) estimating person parameters using weighted maximum likelihood (WML; Warm, 1989),

(iii) computing the response residuals [formula (2)],

(iv) computing the empirical correlation matrix,

(v) extracting the largest value from the correlation matrix.

in order to find the empirical 95th and 99th percentiles. Note that we only simulate data sets under the null hypothesis, there is no local dependence in the simulated data sets.

Results: Simulation Study

Figure 1 reports the empirical 95th and 99th percentiles in the empirical distribution of the maximum residual correlation for dichotomous items. The top panel shows $\mu = 0$ (labeled 'good targeting') and the bottom panel shows $\mu = 2$ (labeled 'bad targeting'). The reason for this labeling is that the average of the item locations (the item difficulties) is zero.

[Figure 1. The empirical 95th and 99th percentiles in the empirical distribution of $Q_{3,max}$ for dichotomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)]

The percentiles decrease as the sample size increases, and they increase with the number of items. The latter finding is hardly surprising in a comparison of the maximum of 45, 105, and 190 item pairs, respectively. However, it is evident that the targeting does not have an impact on the percentiles. Figure 2 reports the empirical 95th and 99th percentiles in the empirical distribution of the maximum residual correlation for polytomous items. Again the top panel labeled 'good targeting' shows $\mu = 0$ the bottom panel labeled 'bad targeting' shows $\mu = 2$.

[Figure 2. The empirical 95th and 99th percentiles in the empirical distribution of $Q_{3,max}$ for polytomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)]

For N=200 some of these percentiles were very large. Again, the percentiles decrease as sample size increases and the mean μ had little impact on the percentiles.

When we considered item pairs individually and computed the empirical distribution Q_3 for selected item pair there was quite a big difference across item pairs and, again, the percentiles decrease as sample size increases while the mean μ had little impact on the percentiles. Comparing

the percentiles in the distribution of the correlation for a single a priori specified item pair shows that percentiles increase with the number of items (results not shown). Thus, the above finding that the percentiles in the distribution of the maximum correlation increase with the number of items is not solely due to the increase in the number of item pairs. Figures 3 and 4 show the empirical distribution of $Q_{3,*}$ for dichotomous and polytomous items, respectively.

[Figure 3. The empirical 95th and 99th percentile in the empirical distribution of $Q_{3,*}$ for dichotomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)]

[Figure 4. The empirical 95th and 99th percentile in the empirical distribution of $Q_{3,*}$ for polytomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)]

When using $Q_{3,*}$ rather than $Q_{3,max}$ there is a smaller effect of the number of items, but again the critical values decrease as sample size increases.

Makransky and Bilenberg data

Methods: Makransky and Bilenberg data

The empirical data example uses the ADHD rating scale (ADHD-RS-IV), which has been validated using the Rasch model in a sample consisting of 566 Danish school children (52% boys), ranging from 6 to 16 years of age (mean = 10.98) by Makransky and Bilenberg (2014). The parent and teacher ADHD-RS-IV (Barkley et al., 1999) which is one of the most frequently-used scales in treatment evaluation of children with ADHD consists of 26 items which measure across three subscales: inattention, hyperactivity/impulsivity and conduct problems. Parents and teachers are independently asked to rate children on the 26 items on a 4-point Likert-type scale, resulting in 6

subscales (three with ratings from parents and three with ratings from teachers). In this study we will specifically focus on the nine items from the teacher ratings of the hyperactivity/impulsivity subscale. We attempted to find the empirical residual correlation critical value that should be applied to indicate LD. We did this by simulating data sets under the Rasch model, i.e. data sets without local dependence. Using an implementation in SAS (Christensen, 2006), the simulation study was conducted by simulating 10,000 data sets under the Rasch model and, for each of these, performing the steps (i)-(v) outlined above in order to find the empirical 95th and 99th percentiles.

Results: Makransky and Bilenberg data

In this section we describe an empirical example where we illustrate the practical challenge of deciding whether or not the evidence of LD provided by the maximum value $Q_{3,max}$ of Yen's (1984) Q_3 is large enough to violate the assumptions of the Rasch model. Makransky and Bilenberg (2014) report misfit to the Rasch model using a critical value of 0.2 to indicate LD. Using this critical value they identified LD between item 2 ("Leaves seat") and item 3 ("Runs about or climbs excessively") where Q_3 was 0.26, and also between item 7 ("Blurts out answers") and item 8 ("Difficulty awaiting turn") where Q_3 was 0.34 (Table 1).

[Table 1. The observed residual correlation matrix in the Makransky and Bilenberg (2014) data for the teacher ratings of Hyperactivity/Impulsivity in the ADHD-RS-IV.]

They were able to explain the LD based on the content of the items, e.g. that students would have to leave their seat in order to run about or climb excessively within a classroom environment, where students are usually required to sit in their seat, and they went on to adjust the scale based on

these results. Thus, the observed value of $Q_{3,max}$ is 0.34, and since the average correlation $\bar{Q}_3$ in Table 1 is -0.12 the observed value of $Q_{3,*}$ is 0.46.

As described above, there are examples in the literature where this procedure has been used with critical values of Q_3 ranging from 0.1 to 0.7. The choice of the critical value has implications for the interpretation of the measurement properties of a scale. This will, in turn, impact upon any amendments that might be made, as well as the conclusions that are drawn. Using a critical value of 0.3 would lead to the conclusion that the residual correlation value of 0.26 identified between items 2 and 3 is not in violation of the Rasch model. A critical value of 0.7 would lead to the conclusion that there is no LD in the scale. Alternatively, a critical value of 0.1 would result in the conclusion that three additional pairs of items also exhibit LD within this data set.

Based on the estimated item and person parameters in the Makransky and Bilenberg data, we simulated 10,000 data sets from a Rasch model without local dependence, computed residuals and their associated correlations. The empirical distribution of the maximum value $Q_{3,max}$ based on these 10,000 data sets is shown in Figure 5.

[Figure 5. The empirical distribution of $Q_{3,max}$ based on 10,000 data sets simulated using item and person parameters from the Makransky and Bilenberg (2014) data.]

The 95th and 99th percentiles in this empirical distribution were 0.19, and 0.24, respectively indicating that Makransky and Bilenberg were correct in concluding that $Q_{3,max}$ =0.34 indicated misfit. Using the parametric bootstrap results reported in Figure 1, Makransky and Bilenberg could have rejected the assumption of no LD with a p-value of p<0.001. For nine items (as in the Makransky and Bilenberg data), there are 36 item pairs, and based on the simulated data sets we are able to determine critical values for Yen's Q_3 for each item pair. If a hypothesis about LD had been specified a priori for a single item pair (e.g. between items 2 and 3), then it would make sense to

compare the observed correlation to a percentile in the empirical distribution of correlations for this item pair. In Table 2 we show the median and four empirical percentiles.

[Table 2. Empirical 95th and 99th percentiles in the empirical distribution of the correlations of residuals. Based on 10,000 data sets simulated under the Rasch model using estimated parameters from the Makransky and Bilenberg (2014) data.]

Table 2 illustrates that the median value of the Q_3 test statistic for any item pair is negative. Table 2 further outlines the critical values that could be used for tests at the 5% and 1% level respectively, if the hypothesis about LD was specified a priori for an item pair. These values ranged from 0.05 to 0.07 with a mean of 0.06 for the 95th, and from 0.09 to 0.14 with a mean of 0.12 for the 99th percentiles. Since no a priori hypotheses about LD were made in the Makransky and Bilenberg study, the results indicate that the conclusions made using a critical value of 0.2 were reasonable. Since the simulation performed is based on the estimated item and person parameters in the Makransky and Bilenberg data it can be viewed as a parametric bootstrap approach.

The empirical distribution of $Q_{3,*}$ (the difference between $Q_{3,\max}$ and the average correlation $\bar{Q}_3$) based on these 10,000 data sets is shown in Figure 6.

[Figure 6. The empirical distribution of $Q_{3,*}$ based on 10,000 data sets simulated using item and person parameters from the Makransky and Bilenberg (2014) data.]

Since the average value $\bar{Q}_3$ is negative it is not surprising that the distribution of the $Q_{3,*}$ is shifted to the right compared to the distribution of $Q_{3,max}$. The relevant critical value for a test at the 5% level is 0.26 and the relevant critical value for a test at the 1% level is 0.31. The observed

value of the average correlation being $\bar{Q}_3 = -0.12$, as computed from Table 1, we see that $Q_{3,*} = 0.46$. Based on this Makransky and Bilenberg were correct in concluding that LD exists in the data.

Formally the results in Figures 1 and 2 would enable us to reject the overall hypothesis about absence of LD and conclude that there is LD for the item pair 7 and 8. Of course a parametric bootstrap approach like this could be extended from looking at the maximum value $Q_{3,max}$ to looking at the empirical distribution of largest and the second largest Q_3 value. Makransky and Bilenberg report that LD between the items was successfully dealt with by combining the item pairs with LD into single combination items, and evaluating fit for the resulting seven item scale. They further argue that item deletion is not desirable because the Hyperactivity/Impulsivity subscale in the ADHD-RS-IV is '… developed to assess the diagnosis in the DSM-IV and DSM-5, and the elimination of the items would decrease the content validity of the scale' (Makransky and Bilenberg, 2014; p.702). A third alternative is to model the LD using log linear Rasch models (Kelderman, 1984). Table 3 outlines the result obtained using item deletion and combining items, respectively. Item fit was evaluated using comparison of observed and expected item-restscore correlation (Kreiner, 2011), while Andersens (1973) conditional likelihood ratio test was used to evaluate scale fit.

[Table 3. Evaluation of item and scale fit in four models with item deletion and a model with item combination the Makransky and Bilenberg data. Item fit evaluated using comparison of observed and expected item-restscore correlations, total fit based on Andersens (1973) conditional likelihood ratio test, P-values rapported.]

Based on the results in Table 3 we see that combining items yields the best item and scale fit. The four models are also compared with respect to the test information function (Figure 7).

[Figure 7. The test information in four models with item deletion and in the model with item combination.]

Figure 7 shows that combining items yields the highest test information.

Discussion

Local independence implies that, having extracted the unidimensional latent variable, there should be no leftover patterns in the residuals (Tennant and Conaghan, 2007). We simulated the distribution of residuals that can be expected between two items when the data fit the Rasch model under a number of different conditions. In all instances, the critical values used to indicate LD were shown to be lower when there are fewer items, and more cases within a dataset. Similar patterns were observed for dichotomous and polytomous items.

In the first part of this study, empirical percentiles were reported from the empirical distribution of the $Q_{3,max}$ test statistic and the $Q_{3,*}$ test statistic. We reported critical values across a number of situations with differing numbers of items, response options, and respondents and with different targeting. Each of these conditions was based on 10,000 data sets simulated under the Rasch model. The outlined parametric bootstrap method could be applied on a case by case basis to inform research about a reasonable choice of cut point for the maximum value of the $Q_{3,max}$ and for the $Q_{3,*}$ test statistics. The second part of this study made it clear that the critical value of the $Q_{3,max}$ test statistic depends heavily on the number of items, but that the $Q_{3,*}$ test statistics are more stable.

In the second part of this study, the empirical 95th and 99th percentiles were reported from the empirical distribution of the maximum value $Q_{3,max}$ of Yen's (1984) Q_3 test statistic in 10,000 data sets, which were simulated under the Rasch model using the estimated item and person parameters

from the Makransky and Bilenberg (2014) data. Based on this, a critical value of 0.19 was observed at the 95th percentile and a critical value of 0.24 was observed at the 99th percentile. Since the observed value was $Q_{3,max} = 0.34$, it is reasonable to conclude that there is LD in the data set.

Having disclosed evidence of LD when it is found to exist, several ways of dealing with it have been suggested. These include the deletion of one of the LD items or by fitting the partial credit model to polytomous items resulting from summation locally dependent Rasch items (Andrich, 1985; Kreiner and Christensen, 2007; Makransky and Bilenberg, 2014). Other approaches include using testlet models (Wilson and Adams, 1995; Wang and Wilson, 2005) or a bi-factor model (Reise, 2012). In our analysis of the Makransky and Bilenberg (2014) data we found that combining items yielded the best item and scale fit and the highest test information.

Summary and Recommendations

In summary, several methods for identifying LD have been suggested, but the most frequently used one appears to be Yen's Q3 based on computing residuals (observed item responses minus their expected values), and then correlating these residuals. Thus, in practice, LD is identified through the observed correlation matrix of residuals based on estimated item and person parameters, and residual correlations above a certain value are used to identify items that appear to be locally dependent.

It was shown that a singular critical value for the $Q_{3,max}$ test statistics is not appropriate for all situations, as the range of residual correlations values is influenced by a number of factors. The critical value which indicates LD will always be relative to the parameters of the specific dataset, and various factors should be considered when assessing LD. For this reason, the recommendation by Marais (2013) was that LD should be considered relative to the average residual correlation and

thus that the $Q_{3,*}$ test statistic should be used. For neither of the test statistics a single stand-alone critical value exists.

Despite no single critical value being appropriate, our simulations show that the $Q_{3,*}$ critical value appears to be reasonably stable around a value of 0.2 above the average correlation. Within the parameter ranges that were tested, any residual correlation >0.2 above the average correlation would appear to indicate LD, and any residual correlation of independent items at a value >0.3 above the average would seem unlikely.

Finch and Jeffers (2015) proposed a permutation test for local dependence based on the Q3 and found it to have good Type I error control, while also yielding more power for detecting LD than the use of the 0.2 cut-value. Bootstrapping and determining critical values for the Q3 is one option, but using one of the statistics with known null distribution listed in the introduction is a better option. For researchers for whom these tests are not available the results presented in Figures 3 and 4 yield guideline for choosing a critical value of the $Q_{3,max}$ and the results presented in Figures 5 and 6 yield guideline for choosing a critical value of the $Q_{3,*}$ for certain data structure situations and the parametric bootstrap approach outlined illustrates how a precise critical value can be ascertained. A complete summary of our simulation studies is available online on the homepage

http://publicifsv.sund.ku.dk/~kach/Q3/critical_values_Yens_Q3.html

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Andrich, D. (1985). A latent trait model for items with response dependencies: Implications for test construction and analysis. In S. E. Embretson (Ed.), Test design (pp. 245-275). New York, NY: Academic Press.

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.

Andrich, D., Humphry, S., & Marais, I. (2012). Quantifying local, response dependence between two polytomous items using the rasch model. Applied Psychological Measurement, 36, 309-324.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. Applied Psychological Measurement, 34, 181-192.

Andrich, D., & Luo, G. (2003). Conditional pairwise estimation in the Rasch model for ordered response categories using principal components. Journal of Applied Measurement, 4, 205-221.

Andrich, D., Sheridan, B., & Luo, G. (2010). RUMM2030 [Computer software and manual]. Perth, Australia.

Barkley, R., Gwenyth, E. H., & Arthur, L. R. (1999). *Defiant teens. A clinician's manual for* assessment and family intervention. New York, NY: Guilford Press.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics, 22, 265-289.

Christensen, K. B. (2006). Fitting polytomous Rasch models in SAS. Journal of Applied
Measurement, 7, 407-417.

Das Nair, R., Moreton, B. J., & Lincoln, N. B. (2011). Rasch analysis of the Nottingham extended
activities of daily living scale. Journal of Rehabilitation Medicine, 43, 944-950.

Davidson, M., Keating, J. L., & Eyres, S. (2004). A low back-specific version of the SF-36 physical
functioning scale. Spine, 29, 586-594.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with
conditional covariance functions. Journal of Educational and Behavioral Statistics, 23, 129-
151.

Finch, W. H., & Jeffers, H. (2016). A Q3-based permutation test for assessing local independence.
Applied Psychological Measurement, 40, 157-160.

Fisher, R.A. (1915). Frequency distribution of the values of the correlation coefficient in samples of
an indefinitely large population. Biometrika, 10, 507-521.

Glas, C. A. W., & Suarez-Falcon, J. C. (2003). A Comparison of item-fit statistics for the three-
parameter logistic model. Applied Psychological Measurent, 27, 87-106.

González-de Paz, L., Kostov, B., López-Pina, J. a, Solans-Julián, P., Navarro-Rubio, M. D., & Sisó-
Almirall, A. (2015). A Rasch analysis of patients' opinions of primary health care professionals'
ethical behaviour with respect to communication issues. Family Practice, 32, 237-243.

Haberman, S. J. (2007). The interaction model. In M. von Davier & C. H. Carstensen (Eds.),
Multivariate and mixture distribution Rasch models: Extensions and applications (pp. 201-
216). New York, NY: Springer.

Hissbach, J. C., Klusmann, D., & Hampe, W. (2011). Dimensionality and predictive validity of the HAM-Nat, a test of natural sciences for medical school admission. BMC Medical Education, 11, 83.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hoskens, M., & De Boeck, P. (1997). A parametric model for local dependence among test items. Psychological Methods, 2, 261-277.

Ip, E. H. (2001). Testing for local dependency in dichotomous and polytomous item response models. Psychometrika, 66, 109-132.

Ip, E. H. (2002). Locally dependent latent trait model and the Dutch identity revisited. Psychometrika, 67, 367-386.

Kelderman, H. (1984). Loglinear Rasch model tests. Psychometrika, 49, 223-245.

Kreiner, S., & Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. Communications in Statistics - Theory and Methods, 33, 1239-1276.

Kreiner, S., & Christensen, K. B. (2007). Validity and Objectivity in Health-Related Scales: Analysis by Graphical Loglinear Rasch Models. In M. von Davier & C. H. Carstensen (Eds.), Multivariate and Mixture Distribution Rasch Models: Extensions and Applications (pp. 329-346). New York: Springer-Verlag.

Kreiner, S., & Christensen, K. B. (2011). Exact evaluation of bias in Rasch Model residuals. In A. R. Baswell (Ed.), Advances in Mathematics Research vol.12 (pp. 19-40). Nova publishers.

Kreiner S. (2011). A Note on Item-Restscore Association in Rasch Models. Applied Psychological Measurement, 35, 557-561.

La Porta, F., Franceschini, M., Caselli, S., Cavallini, P., Susassi, S., & Tennant, A. (2011). Unified balance scale: an activity-based, bed to community, and aetiology-independent measure of balance calibrated with Rasch analysis. Journal of Rehabilitation Medicine, 43, 435-444.

Liu, Y., & Maydeu-Olivares, A. (2013). Local dependence diagnostics in IRT modeling of binary data. Educational and Psychological Measurement, 73, 254-274.

Liu, Y., & Thissen, D. (2012). Identifying local dependence with a score test statistic based on the bifactor logistic model. Applied Psychological Measurement, 36, 670-688.

Liu, Y., & Thissen, D. (2014). Comparing score tests and other local dependence diagnostics for the graded response model. British Journal of Mathematical and Statistical Psychology, 67, 496-513.

Lazarsfeld, P. F., & Henry, N.W. (1968). Latent structure analysis. Boston: Houghton Mifflin.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Lucke, J. F. (2005). "Rassling the Hog": The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. Applied Psychological Measurement, 29, 106-125.

Makransky, G., & Bilenberg, N. (2014). Psychometric properties of the parent and teacher ADHD rating scale (ADHD-RS): Measurement Invariance Across Gender, Age, and Informant. Assessment, 21, 694-705.

Makransky, G., Rogers, M. E., & Creed, P. A. (2014). Analysis of the construct validity and measurement invariance of the career decision self-efficacy scale: A Rasch model approach. Journal of Career Assessment, 21, 694-705.

Marais, I., & Andrich, D. (2008a). Formalising dimension and response violations of local independence in the unidimensional Rasch model. Journal of Applied Measurement, 9, 1-16.

Marais, I., & Andrich, D. (2008b). Effects of varying magnitude and patterns of local dependence in the unidimensional Rasch model. Journal of Applied Measurement, 9, 105-124.

Marais, I. (2009). Response dependence and the measurement of change. Journal of Applied Measurement, 10, 17-29.

Marais, I. (2013). Local dependence. In K. B. Christensen, S. Kreiner, M. Mesbah (Eds.), Rasch models in health (pp. 111-130). Wiley-ISTE Ltd., London, UK.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149–174.

Maydeu-Olivares, A., & Liu, Y. (2015). Item diagnostics in multivariate discrete data. Psychological Methods, 20, 276-292.

Ramp, M., Khan, F., Misajon, R. A., & Pallant, J. F. (2009). Rasch analysis of the Multiple Sclerosis Impact Scale MSIS-29. Health and Quality of Life Outcomes, 7, 58.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J. A., … Hambleton, R. K. (2007). Psychometric evaluation and calibration of health-related quality of life item banks. Medical Care, 45(Suppl 1), S22-S31.

Reise, S. P. (2012). The rediscovery of bifactor measurement models, Multivariate Behavioral Research, 47, 667-696.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Røe, C., Damsgård, E., Fors, T., & Anke, A.. (2014). Psychometric properties of the pain stages of change questionnaire as evaluated by Rasch analysis in patients with chronic musculoskeletal pain. BMC Musculoskeletal Disorders.15, 95.

Scott, S. L., & Ip, E. H. (2002). Empirical Bayes and item-clustering effects in a latent variable hierarchical model. Journal of the American Statistical Association, 97, 409-419.

ten Klooster, P. M., Taal, E., & van de Laar, M. A. F. J. (2008). Rasch analysis of the Dutch health assessment questionnaire disability index and the health assessment questionnaire II in patients with rheumatoid arthritis. Arthritis & Rheumatism, 59: 1721-1728.

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care and Research. 57, 1358-1362.

van den Wollenberg, A. L. (1982). Two new test statistics for the Rasch model. Psychometrika, 47, 123-139.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence, Journal of Educational Measurement, 30, 187-213.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. Applied Psychological Measurement. 29, 126-149.

Warm, T. (1989). Weighted likelihood estimation of ability in item response theory. Psychometrika, 54, 427-450.

Wilson, M., & Adams, R. J. (1995). Rasch models for item bundles. Psychometrika, 60, 181-198,

Zwinderman, A. H. (1995). Pairwise parameter estimation in Rasch models. Applied Psychological Measurement, 19, 369-375.

Figure 1. The empirical 95th and 99th percentiles in the empirical distribution of $Q_{3,max}$ for dichotomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)
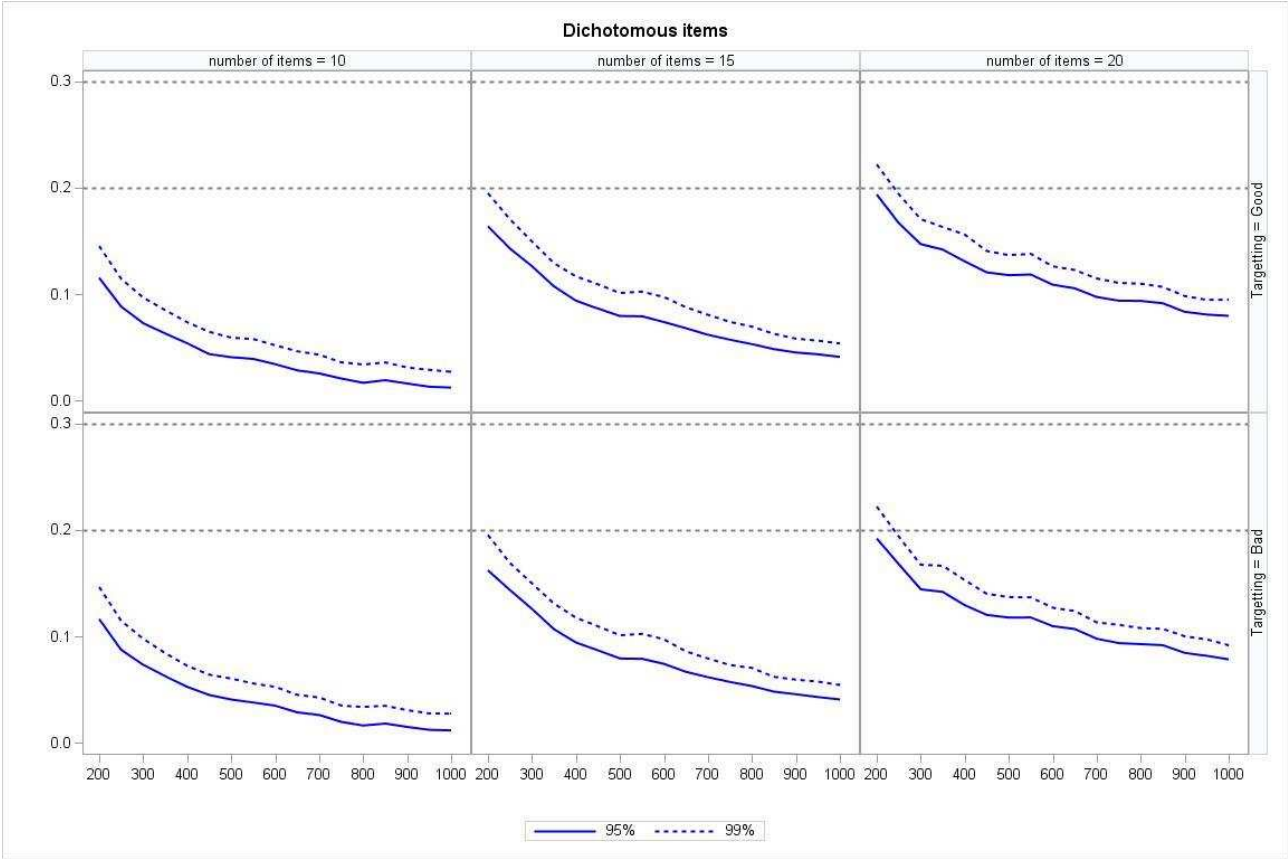
Figure 2. The empirical 95th and 99th percentiles in the empirical distribution of $Q_{3,max}$ for polytomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)
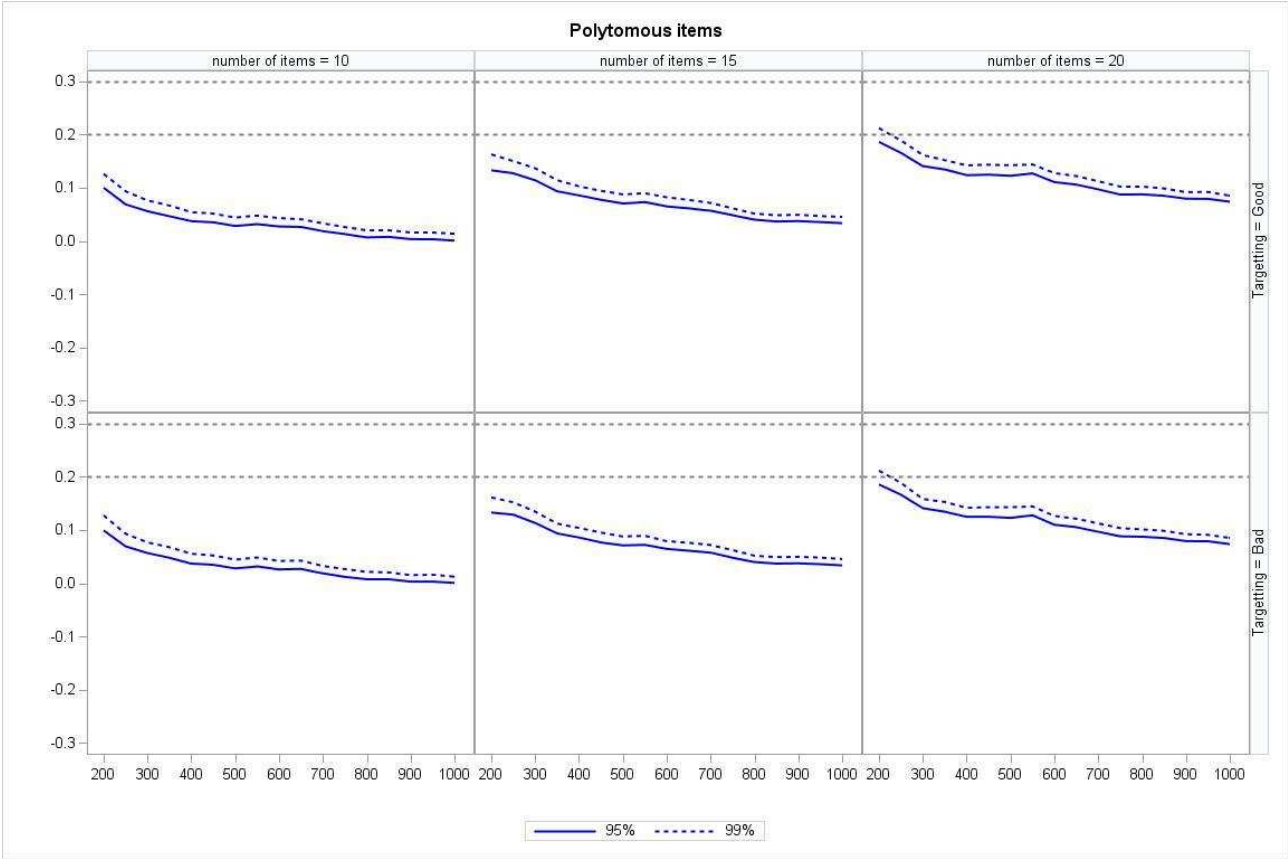
Figure 3. The empirical 95th and 99th percentiles in the empirical distribution of $Q_{3,*}$ for

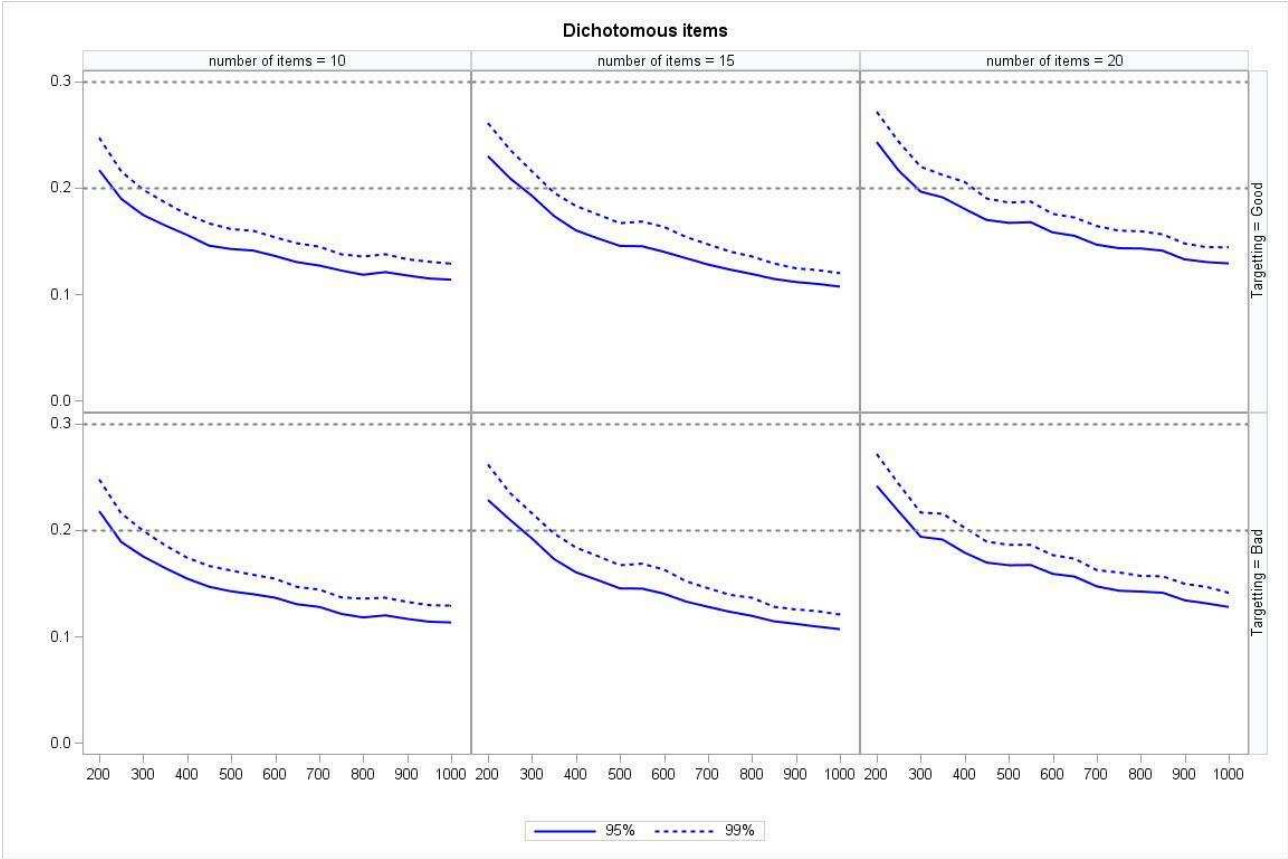dichotomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)

Figure 4. The empirical 95th and 99th percentiles in the empirical distribution of $Q_{3,*}$ polytomous items (grey horizontal dashed lines indicate 0.2 and 0.3, respectively)
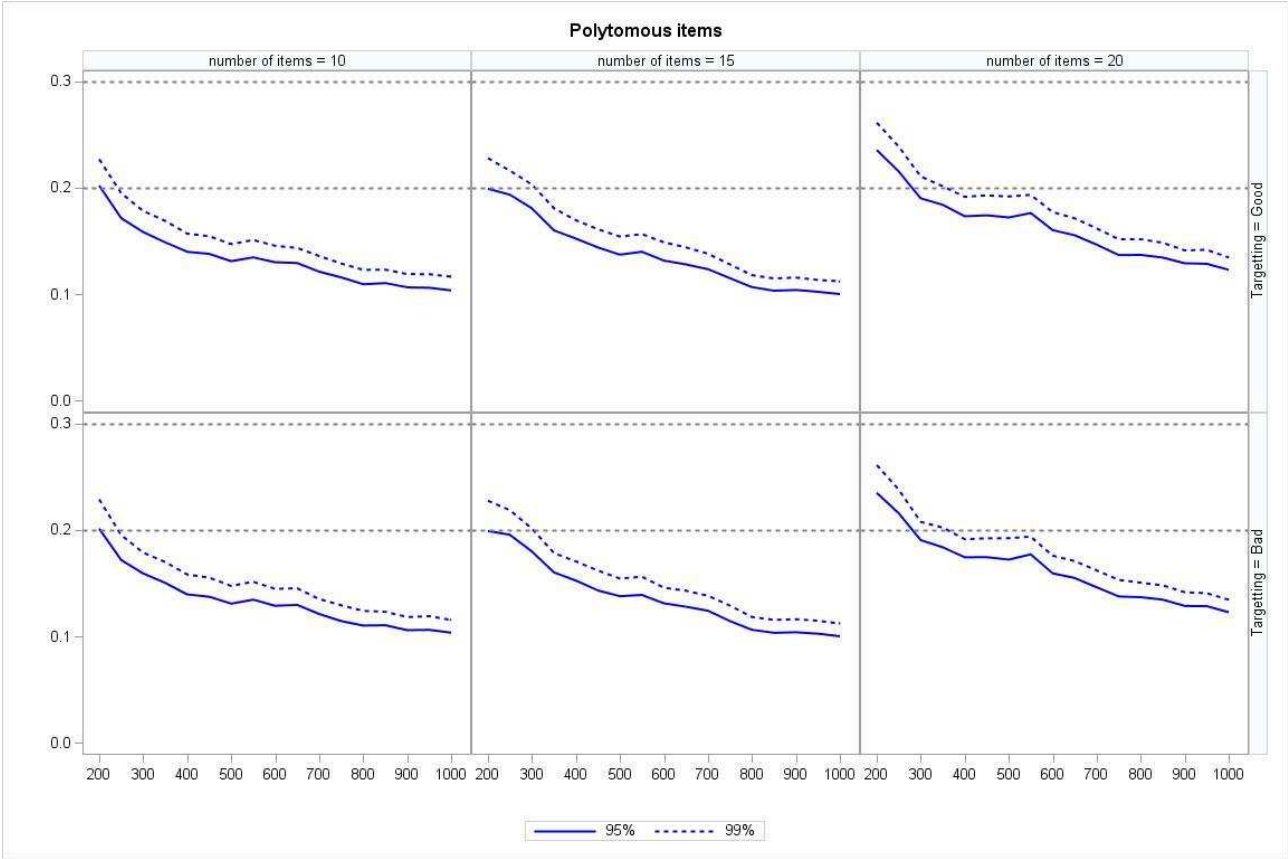
Figure 5. The empirical distribution of the maximum correlation $Q_{3,max}$ between residuals based on 10000 data sets simulated using item and person parameters from the Makransky and Bilenberg (2014) data.
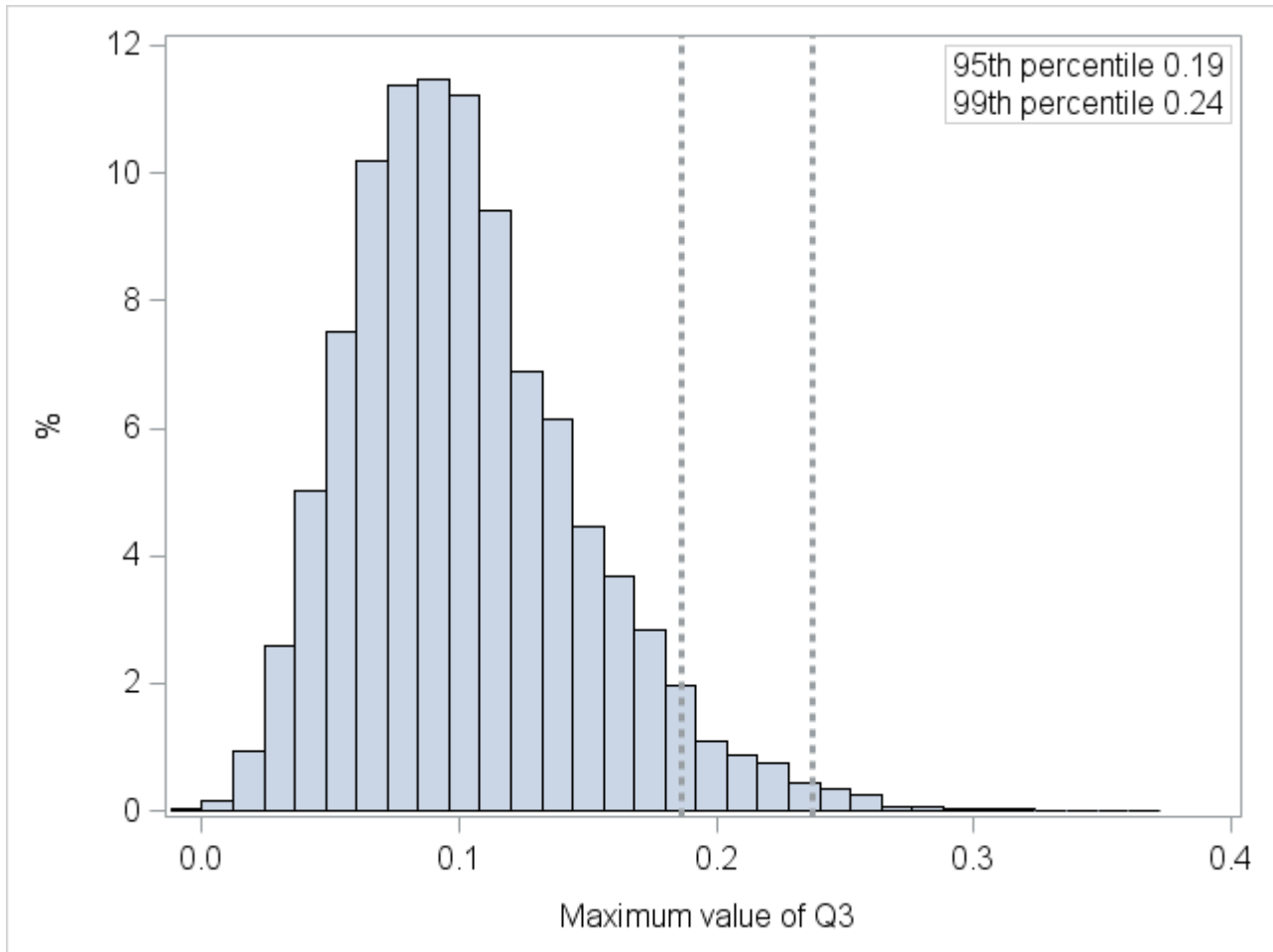
Figure 6. The empirical distribution of the $Q_{3,*}$ test statistic (difference between $Q_{3,max}$ and the average correlation) based on 10000 data sets simulated using item and person parameters from the Makransky and Bilenberg (2014) data.
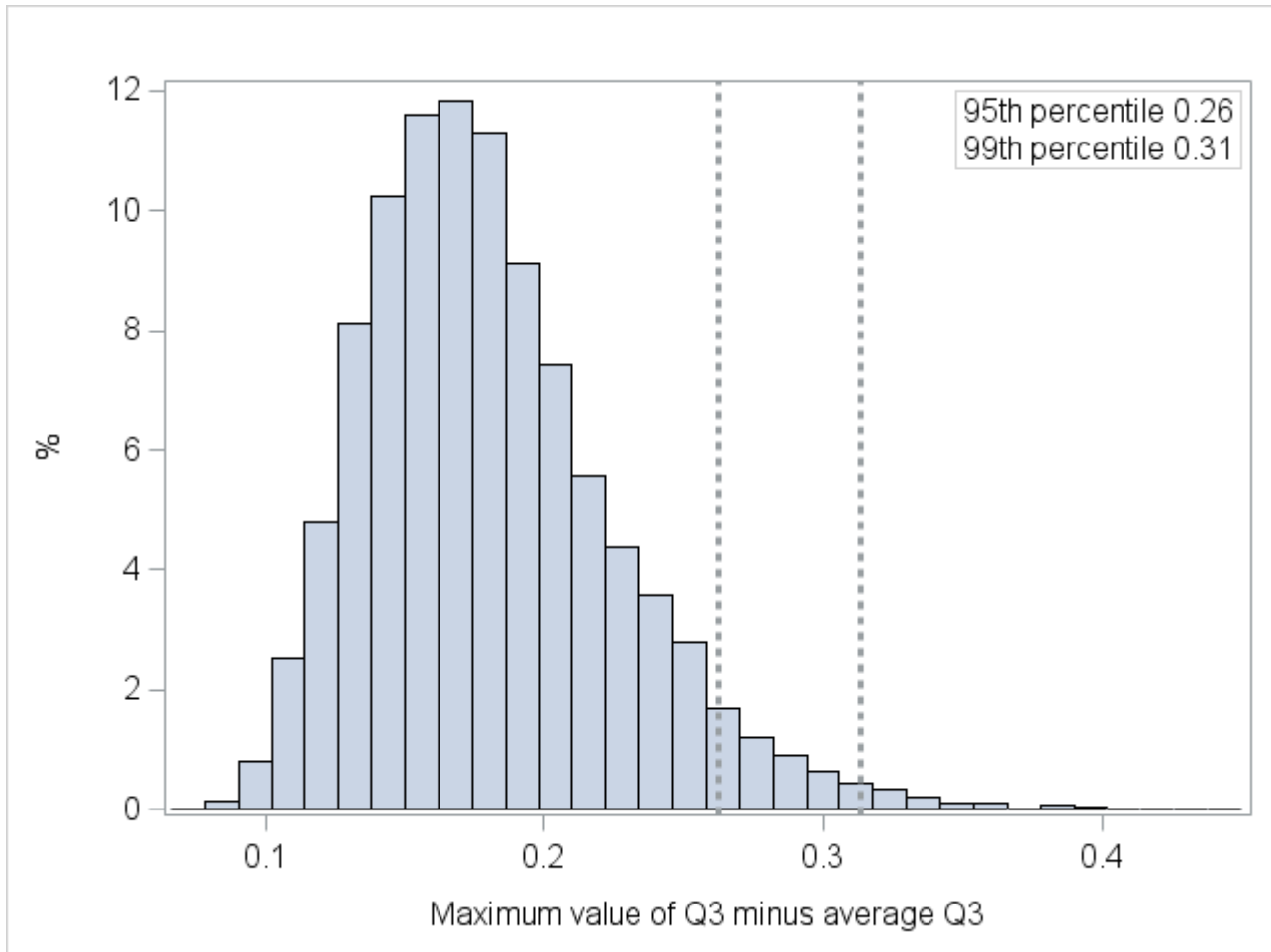
Figure 7. The test information in four models with item deletion and in the model with item combination.
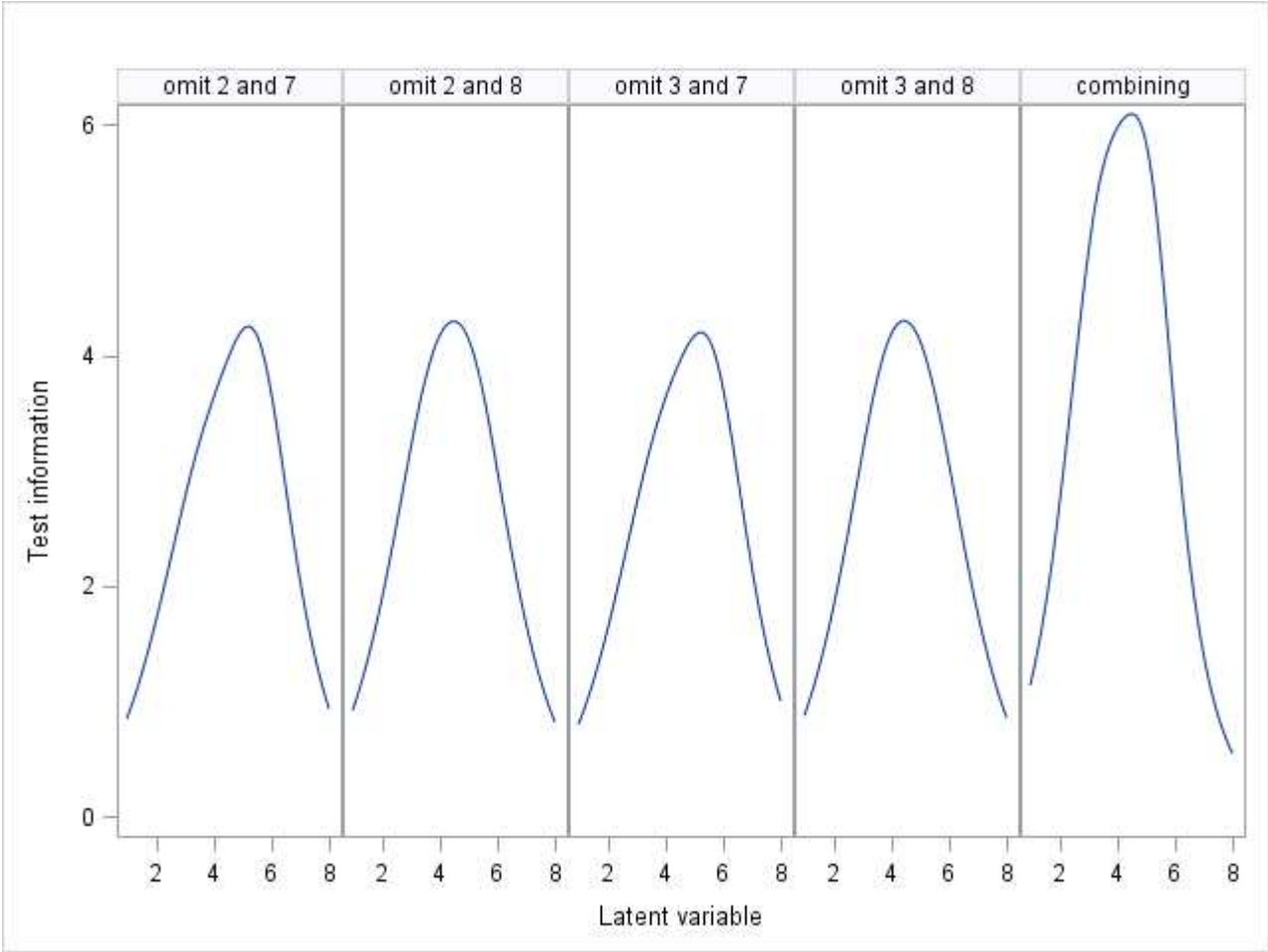
Table 1. The observed residual correlation matrix in the Makransky and Bilenberg data for the

teacher ratings of Hyperactivity/Impulsivity in the ADHD-RS.

Item	1	2	3	4	5	6	7	8	9
1 (Fidgets or squirms)	1.00								
2 (Leaves seat)	**0.12**	1.00							
3 (Runs about or climbs excessively)	0.03	**0.26**	1.00						
4 (Difficulty playing quietly)	-0.05	-0.04	0.04	1.00					
5 (On the go)	-0.09	-0.25	-0.02	-0.14	1.00				
6 (Talks excessively)	-0.20	-0.25	-0.26	-0.21	0.03	1.00			
7 (Blurts out answers)	-0.34	-0.26	-0.23	-0.25	-0.18	0.00	1.00		
8 (Difficulty awaiting turn)	-0.29	-0.21	-0.23	-0.24	-0.19	-0.12	**0.34**	1.00	
9 (Interrupts)	-0.20	-0.12	-0.14	-0.14	-0.24	-0.32	**0.12**	**0.12**	1.00

Table 2. The empirical median, 25th, 75th, 95th, and 99th percentiles in the empirical distribution of the correlations of $Q_{3,ij}$ for all item pairs. Based on 10.000 data sets simulated under the Rasch model using estimated parameters from the Makransky and Bilenberg data.

Item1	item2	Median	(IQR)	Percentile	
				95th	99th
1	2	-0.07	(-0.11 to -0.02)	0.06	0.12
	3	-0.06	(-0.11 to -0.02)	0.06	0.12
	4	-0.08	(-0.13 to -0.02)	0.06	0.12
	5	-0.07	(-0.12 to -0.02)	0.06	0.13
	6	-0.07	(-0.11 to -0.02)	0.07	0.14
	7	-0.07	(-0.12 to -0.02)	0.06	0.12
	8	-0.07	(-0.12 to -0.02)	0.06	0.13
	9	-0.07	(-0.12 to -0.02)	0.06	0.13
2	3	-0.08	(-0.12 to -0.03)	0.05	0.10
	4	-0.09	(-0.15 to -0.04)	0.05	0.10
	5	-0.08	(-0.13 to -0.03)	0.06	0.13
	6	-0.08	(-0.13 to -0.03)	0.06	0.12
	7	-0.08	(-0.13 to -0.03)	0.05	0.12
	8	-0.08	(-0.12 to -0.03)	0.05	0.12
	9	-0.08	(-0.13 to -0.03)	0.05	0.12
3	4	-0.10	(-0.15 to -0.05)	0.03	0.09
	5	-0.08	(-0.13 to -0.03)	0.05	0.11
	6	-0.07	(-0.13 to -0.02)	0.06	0.12
	7	-0.08	(-0.13 to -0.03)	0.05	0.11
	8	-0.08	(-0.13 to -0.03)	0.05	0.11
	9	-0.08	(-0.13 to -0.03)	0.05	0.11
4	5	-0.09	(-0.15 to -0.04)	0.05	0.12
	6	-0.08	(-0.14 to -0.03)	0.06	0.12
	7	-0.10	(-0.15 to -0.04)	0.04	0.11
	8	-0.09	(-0.15 to -0.04)	0.05	0.10
	9	-0.09	(-0.15 to -0.04)	0.05	0.12
5	6	-0.08	(-0.13 to -0.02)	0.06	0.13
	7	-0.08	(-0.13 to -0.03)	0.06	0.12
	8	-0.08	(-0.13 to -0.03)	0.06	0.13
	9	-0.09	(-0.14 to -0.03)	0.06	0.13
6	7	-0.08	(-0.13 to -0.02)	0.06	0.13
	8	-0.08	(-0.13 to -0.02)	0.06	0.13
	9	-0.08	(-0.13 to -0.02)	0.06	0.13
7	8	-0.08	(-0.13 to -0.03)	0.06	0.12
	9	-0.08	(-0.13 to -0.03)	0.06	0.12
8	9	-0.08	(-0.13 to -0.03)	0.06	0.12

Table 3. Evaluation of item and scale fit in four models with item deletion and a model with item combination the Makransky and Bilenberg data. Item fit evaluated using comparison of observed and expected item-restscore correlations, total fit based on Andersens (1973) conditional likelihood ratio test, P-values rapported.

		Original scale	Deleting items				Combining Items
			2 and 7	3 and 7	2 and 8	3 and 8	
Item fit*	1 (Fidgets or squirms)	0.981	0.880	0.251	0.697	0.428	0.297
	2 (Leaves seat)	0.989		0.558		0.695	0.650
	3 (Runs about or climbs excessively)	**0.005**	**0.008**		**0.009**		
	4 (Difficulty playing quietly)	**0.001**	**0.010**	**0.016**	**0.005**	**0.010**	**0.037**
	5 (On the go)	0.716	0.123	0.208	0.180	0.319	0.168
	6 (Talks excessively)	**0.030**	0.124	0.171	0.108	0.150	0.345
	7 (Blurts out answers)	**0.023**			0.116	0.095	0.735
	8 (Difficulty awaiting turn)	0.394	0.761	0.613			
	9 (Interrupts)	0.772	0.996	0.782	0.854	0.906	0.196
Scale Fit**		**0.036**	**0.019**	0.218	**0.002**	0.045	0.081