International Journal of Statistics and Probability; Vol. 8, No. 5; September 2019
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

On the Contextual Conditions Driving a Difficulty Factor

Karl Schweizer[1], & Siegbert Reiß[1]

[1] Institute of Psychology, Goethe University Frankfurt, Frankfurt, Germany

Correspondence: Karl Schweizer, Institute of Psychology, Goethe University Frankfurt, Frankfurt, 60323, Germany. Tel: 49-69-798-353556. E-mail: k.schweizer@psych.uni-frankfurt.de

Received: June 13, 2019 Accepted: July 15, 2019 Online Published: July 29, 2019

doi:10.5539/ijsp.v8n5p1 URL: https://doi.org/10.5539/ijsp.v8n5p1

Abstract

This paper reports three simulation studies conducted to identify the contextual conditions leading to the observation of a difficulty factor in confirmatory factor analysis. The data of each study were generated to show one underlying source of responding only whereas the difficulties of the simulated items constituting the contextual condition were varied. The first study showed that a broad range of difficulties of items was insufficient for driving a difficulty factor. The second study revealed that very large and small difficulties of the same size could lead to a difficulty factor if the confirmatory factor model included two correlated factors. In the third study a subgroup of simulated items showed very large difficulties of the same size while the difficulties of the other simulated item were varied. In this study almost all combinations of difficulties led to the observation of a difficulty factor that was correlated or uncorrelated with the genuine factor.

Keywords: difficulty factor, simulation, binary data, difficulty, variance scaling

1. Introduction

1.1 The Difficulty Factor as a Research Problem

The difficulty factor is described as a factor with factor loadings related to the difficulties of items (Ferguson, 1941). It is usually observed in addition to another factor that is associated with the latent source of responding as, for example, an ability or trait. Following McDonald (1965), we characterize the factor reflecting the latent source as *genuine* factor of content. Whereas the latent source underlying the data is thought to drive the genuine factor, the difficulty factor is considered as a phenomenon that associates with the difficulties of items. If the data require this difficulty factor for achieving good model fit, the typical model of measurement of confirmatory factor analysis is insufficient since it includes one factor only (Graham, 2006). This factor that is also referred to as latent variable can only represent the latent source of responding whereas the source of the neglected difficulty factor causes model misfit. Advanced confirmatory factor models as, for example, the models employed in multitrait-multimethod research (Byrne, 2016) allow for a second factor. Such a factor can capture the variation that is thought to lead to a difficulty factor. However, the additional factor need not to be a difficulty factor, it can also be a substantial factor. Neither the conditions driving a difficulty factor nor the features characterizing a difficulty factor appear to be especially clear.

The research work reported in this paper investigates the condition driving the difficulty factor in the framework of confirmatory factor analysis in order to improve the identifiability of a difficulty factor. Special emphasis is given to the role of the similarity between very easy and very difficult items. Furthermore, the possibility of identifying a specific factor as difficulty factor is investigated.

1.2 Explanations of the Difficulty Factor

The interest in the difficulty factor as research topic appears to have been stimulated by an unusual finding reported by Guilford (1941). In investigating the factorial structure of the Seashore Test of Pitch Discrimination Guilford observed a factor that appeared to reflect the difficulties of the items instead of providing the basis for a conclusive interpretation.

Ferguson (1941) proposed contradictory aims of test construction as reason for the observation of this factor. They were consistency and discrimination that demanded for sets of items that were homogeneous respectively heterogeneous according to difficulty. According to Ferguson a homogeneous set of items could be expected to give rise to a correlation or covariance matrix with a low rank (ideally the rank of one). If this matrix was input to factor analysis, it was unlikely that another factor emerged besides the genuine factor representing the content of the scale. In contrast, discrimination required heterogeneity of the difficulties of the items to assure discriminability in various ranges of an ability or trait dimension. In the context of assessment with binary items different difficulties could be expected to lead to different upper limit for the correlations among items. As a consequence of this variability, the rank the corresponding correlation or covariance matrix would be larger than one. A rank larger than one meant at least one additional factor besides the genuine factor. According to Ferguson "[the] second factor … represents a difference in difficulty between a particular test and the average difficulty of all tests" (p. 326).

Later on Gibson (1959, 1960) pointed out that difficulty factors originated from mutual non-linear regressions between tests, and McDonald (1965) proposed non-linear factor analysis for capturing the difficulty factor. The model for non-linear factor analysis comprised linear and non-linear components for the *genuine* factors of content and *spurious* factors, respectively. The difficulty factor was conceptualized as spurious factor and represented by the quadratic component of the polynomial. In a subsequent publication McDonald and Ahlawat (1974) provided an example of such a difficulty factor that was assumed to originate from non-linear item characteristics. According to the authors the observation of a spurious factor should only by possible if the analysis of the data was conducted by a model that included a non-linear component. This remark even called Guilford's (1941) observation of a difficulty factor into question since non-linear factors analysis was not available at the time of the first observation of a difficulty factor.

The use of special statistical procedures or parceling items is recommended for dealing with the source leading to the difficulty factors (Floyd & Widaman, 1995). The special statistical procedures for investigating dichotomous data appear to have to far not found wide-spread acceptance. The presently preferred way of investigating dichotomous data comprises tetrachoric correlations in combination with a linear model of measurement (Muthen, 1984). The mentioned parceling procedure requires the computation of subscores for small groups of items. If these groups include items showing markedly different difficulties, most of the differences due to different difficulties are averaged out (Kishton & Widaman, 1994). This means the elimination of what is supposed to drive a difficulty factor.

Bandalos and Gerster (2016) introduce a new perspective into the discussion on the source of the difficulty factor. They highlight the similarity among the difficulties of items. This perspective suggests that it is not the broad range of item difficulties per se but the existence of subsets of items showing the same degree of difficulty that drives a difficulty factor. Such a subset causes an inhomogeneity in the sense that the items of this subset are very homogeneous whereas the remaining items show a not so high degree of homogeneity. Such homogeneity of a few very difficult items is likely to give rise to an additional factor in factor analysis, and the factor loadings on this factor may show a pattern that can be interpreted in favor of a difficulty factor: large factor loadings of the items of the subset of difficult items and small factor loadings of the other items.

Finally, it needs to be mentioned that there is another source that can lead to an additional factor besides the genuine factor of content and that may be confused with the difficulty factor. It is the factor originating from the effect of the item positions. Although it was shown that there is no correspondence of the difficulty factor and the factor due to the item-position effect (Schweizer & Troche, 2018; Zeller, Reiss, & Schweizer, 2017), it is possible that there is a confusion of these factors and that the factor due to the item-position effect contributes to the observation of a difficulty factor.

In sum, the difficulty factor is a factor with factor loadings showing sizes that are related to the difficulties of the items while a conclusive interpretation is not achievable. This difficulty factor is a specific factor in the sense that there are only a few high factor loadings associated with high difficulties. It is observed in investigating sets of dichotomous items showing a rather broad range of difficulties of items. Furthermore, similarity of the difficulties of items is considered as source driving the difficulty factor in dichotomous data.

1.3 Specificities of the Method for Investigating the Difficulty Factor

There is agreement that the difficulty factor is associated with a specific pattern of difficulties of items (e.g. Bandalos & Gerster, 2016; Ferguson, 1941). Therefore, the key to the identification of what exactly drives the difficulty factor is the manipulation of the difficulties in a simulation study. The difficulties of items can even be manipulated independently of the basic structure of simulated data. This property enables to rule out that an additional factor is rooted in the basic structure of the data if there is a basic structure that gives rise to one latent variable only and is kept constant. The manipulation of the difficulties can be achieved by computing parcels (Kishton & Widaman, 1994) or dichotomization (Schweizer & Troche, 2018). Whereas the manipulation by parceling is restricted to achievable means, there is no such restriction to the method of dichotomization.

Data showing the expected basic structure can be generated by a procedure described by Jöreskog and Sörbom (2001). The expected basic structure is induced by the $p \times p$ relational pattern $\mathbf{P}$. This pattern includes a completely uniform matrix as one component and a diagonal matrix with uniform entries in the main diagonal as another component. The uniform matrix is achieved by the multiplication of uniform $p \times 1$ column vectors $\boldsymbol{\lambda}_c$ with the uniform $p \times 1$ row vectors $\boldsymbol{\lambda}_c^T$ and the $p \times p$ diagonal matrix $\mathbf{U}$ by assigning the same value u ($0 < u$) to all entries of the main diagonal $u_{ii} = u$ ($i = 1, \ldots, p$):

$$\mathbf{P} = \boldsymbol{\lambda}_c \boldsymbol{\lambda}_c^T + \mathbf{U} \quad . \tag{1}$$

The vector $\boldsymbol{\lambda}_c$ can be perceived as vector of factor loadings. Uniformity is achieved by selecting the same constant number c ($0 < c < 1$) for all entries:

$$\boldsymbol{\lambda}_c = \begin{bmatrix} c \\ c \\ . \\ . \\ c \end{bmatrix} \quad . \tag{2}$$

This way of generating data creates characteristics that are in line with the tau-equivalent model of measurement (Lord & Novick, 1968):

$$\mathbf{x} = \boldsymbol{\tau}(\xi) + \boldsymbol{\varepsilon} \tag{3}$$

where the $p\times1$ vector $\mathbf{x}$ represents the centered data, the $p\times1$ vector $\boldsymbol{\tau}$ the true components and the $p\times1$ vector $\boldsymbol{\varepsilon}$ the error components. The latent variable ξ that is added in parentheses is no explicit part of the original tau-equivalent model. We added it because is reflects the latent source of responding that is the same for all items. A characteristic of this model is that the entries of $\boldsymbol{\tau}$ correspond to each other:

$$\boldsymbol{\tau} = \begin{bmatrix} \tau \\ \tau \\ . \\ . \\ \tau \end{bmatrix} \tag{4}$$

with $0 < \tau$. For considering this model in combination with confirmatory factor analysis, its formal structure has to be adapted to the structure of the congeneric model of measurement (Jöreskog, 1971) that is nowadays the standard model of confirmatory factor analysis:

$$\mathbf{x} = \boldsymbol{\lambda}_c \xi_c + \boldsymbol{\delta} \quad . \tag{5}$$

The adaptation requires that the component $\boldsymbol{\tau}(\xi)$ is replaced by the product of the $p\times1$ vector $\boldsymbol{\lambda}$ of factor loadings and ξ without parentheses. We employ the subscript c to signify that the equality assumption is made ($\lambda_i = \lambda_j$, $i,j = 1, \ldots, p$) that characterizes the tau-equivalent model. Furthermore, $\boldsymbol{\delta}$ replaces $\boldsymbol{\varepsilon}$. In the absence of contextual sources influencing the structure of data, this model can be expected to provide a good account of data originating from a relational pattern according to the first equation.

In order to find out about the source driving the difficulty factor, conditions need to be established that may lead respectively not lead to a difficulty factor according to what is already known about this factor. The literature on the difficulty factor suggests types of patterns of difficulties characterizing dichotomous data as candidates. Such patterns are achievable by dichotomizing continuous and normally distributed data. Dichotomization means that the $n\times p$ matrix $\mathbf{M}_{\text{continuous}}$ including continuous and normally distributed data is transformed into the $n\times p$ matrix $\mathbf{M}_{\text{dichotomous}}$ of dichotomous and binomially distributed data. Dichotomization can be perceived as function dichot ( ) that accomplishes the transformation:

$$\mathbf{M}_{\text{dichotomous}} = \text{dichot}(\mathbf{M}_{\text{continuous}}) \quad . \tag{6}$$

This function can be defined to yield either zero or one for each element $m_{ij,\text{continuous}}$ of $\mathbf{M}_{\text{continuous}}$ with respect to the cutoff α_j that is specific for column j ($j = 1, \ldots, p$):

$$\text{dichot}(m_{ij,\text{continuous}}) = \begin{cases} 0 \text{ for } m_{ij,\text{continuous}} < \alpha_j \\ 1 \text{ for } m_{ij,\text{continuous}} \geq \alpha_j \end{cases} \quad . \tag{7}$$

The cutoffs need to be selected to achieve specific patterns of difficulty. A broad range of difficulties (Ferguson, 1941; McDonald & Ahlawat, 1974) can be realized by selecting a broad range of cutoffs. The high degree of similarity between very difficult and very easy items (Bandalos & Gerster, 2016) is achievable by the same high and low cutoffs for subsets of items.

If a specific difficulty condition, i.e. a specific pattern of difficulties of items, creates a difficulty factor, this should be apparent as model misfit in confirmatory factor analysis by the one-factor model. In contrast, the two-factor confirmatory factor model with one factor specified to reflect the pattern of difficulties should lead to good model fit. The formal description of this model is given by

$$\mathbf{x} = \boldsymbol{\lambda}_{\mathrm{c}}\xi_{\mathrm{c}} + \boldsymbol{\lambda}_{\mathrm{d}}\xi_{\mathrm{d}} + \boldsymbol{\delta} \quad . \tag{8}$$

This model can be perceived as extension of the one-factor model. It comprises the additional component $\boldsymbol{\lambda}_{\mathrm{d}}\xi_{\mathrm{d}}$ for representing the contribution of the difficulty factor to the response. The $p\times1$ vector $\boldsymbol{\lambda}_{\mathrm{d}}$ of factor loadings and the corresponding latent variable ξ_{d} constitute this component. Both the vector and the latent variable are identified by the subscript *d* to make the association with the difficulty factor obvious.

$$\boldsymbol{\lambda}_{\mathrm{d}} = \begin{bmatrix} d_{\alpha_1} \\ d_{\alpha_2} \\ . \\ . \\ d_{\alpha p} \end{bmatrix} \quad . \tag{9}$$

When fixing factor loadings this way, the exact sizes of the entries are not important but the relationships among the assigned numbers must be accurate. Deviations from the exact size are compensated by the estimation of the variance parameter ϕ_{d} (Schweizer, Troche, & DiStefano, 2019). If the relationships are accurate, the estimation of ϕ_{d} is sufficient for achieving a good account of the data. Furthermore, the size of the estimate of the variance parameter ϕ_{d} signifies whether the factor accounts for a substantial or a negligible amount of variance.

Another important aspect of the model is the specification of the relationship among the latent variables. There is the possibility to assume that the sources underlying the genuine factor and the additional factor are independent of each other. In this case the covariance parameter ϕ_{cd} needs to be specified to zero,

$$\phi_{\mathrm{cd}} = 0 \quad . \tag{10}$$

This possibility excludes that the additional factor also taps the latent source of the genuine factor. The other possibility is estimating the covariance parameter. Selecting this possibility means that the estimation of the variance parameter can yield any value from the set of real numbers, $\phi_{\mathrm{cd}} \in \Re$, but only positive numbers signify a valid result since variances have to be positive.

1.4 The Variance Estimates as Sources of Information

Confirmatory factor analysis as method for investigating the structural validity of scales relies on the outcome of investigating model fit (Deng, Yang, & Marcoulides, 2018). If the model of measurement fits to the structure of the data, good model fit can be expected. Since such investigations are mostly conducted on the basis of the congeneric model of measurement (Jöreskog, 1971), data originating from a uniform relational pattern are likely to yield good model fit whereas inhomogeneity can be expected to lead to model misfit. The configuration of difficulties described by Bandalos and Gerster (2016) can be perceived as such inhomogeneity. One way of dealing with such inhomogeneity is replacing the one-factor model of measurement by a two-factor model. If the two-factor model leads to good model fit, this result implicitly signifies that there is additional variation, which is not captured by the first factor but by the second one.

However, the source of the variation captured by the second factor is not completely clear. There is the possibility of additional variation that is due to the difficulties of the items. This possibility reflects Ferguson's (1941) argument that heterogeneity of difficulties increases the rank of a matrix. Another possibility is that the variation due to the latent source is split up into two parts and each factor captures one of these parts. A further possibility could be that the factors account for the same variance. Allowing the factors to correlate with each other enables this possibility. In order to find out about the nature of the variation that is captured by the factors, the variance estimates ϕ_{c} and ϕ_{d} need to be investigated. Furthermore, it is necessary to estimate the whole variance at the latent level. We conceive the *whole variance* as variance of the sum of the two latent variables var($\xi_{\mathrm{c}} + \xi_{\mathrm{d}}$):

$$\mathrm{var}(\xi_\mathrm{c} + \xi_\mathrm{d}) = \mathrm{var}(\xi_\mathrm{c}) + \mathrm{var}(\xi_\mathrm{d}) + 2r_\mathrm{cd} SD_\mathrm{c} SD_\mathrm{d} \quad . \tag{11}$$

Since the available scaling methods for the variances of latent variables lead to different estimates, it is necessary to assure comparability of the variances by the selection of an appropriate scaling method (Schweizer, Troche, & Distefao, 2019). Criterion-based scaling with squared factor loadings can establish comparability. If the criterion number of these methods is set equal to one, reasonably large variance estimates are achievable (Schweizer, Troche, & Reiß, 2017).

1.5 Objectives

The first objective was to find out whether a broad range of difficulties of items impaired model fit in comparison to difficulties of the same size when investigated by the one-factor model. In addition it was to be investigated whether including an additional factor with factor loadings according to the difficulties of the items improved model fit. The second objective focused on high and low difficulties of items of the same size. It was to be investigated whether subsets of items showing the same high or low difficulties led to a worse model fit than items showing the same medium difficulty. Besides the one-factor model a two-factor model was to be employed. The factor loadings of the second factor of the two-factor model had to reflect the factor loadings. The third objective highlighted the combination of similarity and very large sizes of the difficulties of one subset of items. This investigation was expected to reveal whether having one subset of item with the same extremely high difficulty impaired model fit. The fourth aim was to explore whether and how the variance estimates of the factors changed as a function of the difficulties of the items and the characteristics of the model. Further insight into what stimulated the difficulty factor was expected from this objective.

2. The Effect of a Broad Range of Difficulties of Items - Study 1

2.1 Introduction

Since this study was designed according to the first objective, it was necessary to generate data according to one underling source of responding in combination with different ranges of the difficulties of items. Following the reasoning by Ferguson (1941) and McDonald and Ahlawat (1974), it was expected that increasing the range of the difficulties of items would impair model fit when the investigation was conducted by the one-factor confirmatory factor model but not when using the two-factor model.

2.2 Method

Data were generated according to one latent source of responding, as is described in the theoretical section. The factor loadings for the generation of the relational pattern were specified to 0.4 and the unique variances were set to 0.84. Two-hundred matrices were generated as recommended by Bandalos and Gagné (2012) for simulation studies. We considered the columns of the generated 500 × 20 matrices as simulated items and the rows as simulated test takers. Subsequently, the continuous and normally distributed data were transformed into dichotomous data in such a way that the difficulties linearly increased from the lower to upper limits. These limits were 0.4 and 0.6, 0.3 and 0.7, 0.2 and 0.8, 0.1 and 0.9. Furthermore, equal mean difficulties (0.5-0.5) were considered; they served as comparison level.

Probability-based covariances were computed and used as input to confirmatory factor analysis. Furthermore, for overcoming the difference between the binomially distributed data and the latent variables following the normal distribution, a link transformation was performed in modeling the data (McCullagh & Nelder, 1985; Schweizer, 2013; Schweizer, Ren, & Wang, 2015). Otherwise tetrachoric correlations would have to be computed. However, tetrachoric correlations frequently led to matrices that were not positive definite. Although this could be prevented by the stimulation of the ridge option, the call of this option was avoided since it implicitly diminished the correlations among items.

One-factor and two-factor confirmatory factor models were employed for investigating the data. The factor loadings were specified, as is described in the theoretical section. Since it could be assumed that the latent source contributed equally to each item, the factor loadings on the first latent variable were specified to show the same size; the number one was selected for this purpose. The second latent variable was designed as difficulty factor. The factor loadings were specified to correspond to the difficulties selected for the construction of the data. The variance parameters of the latent variables were set free for estimation to enable the fitting of the model to the data.

The simulation was conducted in several steps. The structured random data were generated by means of PRELIS (Jöreskog & Sörbom, 1999). Additional transformations of the data (e.g. the dichotomization of the continuous data) were conducted by PASCAL programs.

The maximum likelihood estimation method of LISREL (Jöreskog & Sörbom, 2006) was employed for investigating the data. The evaluation of the results was conducted by the following fit indices and criteria (in parenthesis): χ^2,

RMSEA (≤ .06), SRMR (≤ .08), CFI (≥ .95), TLI (≥ .95), and AIC (see DiStefano, 2016; Hu & Bentler, 1999). The CFI difference with a cutoff of .01 served the comparison of models (Cheung & Rensvold, 2002). The chi-square difference test was also considered when it was applicable. The variances of the latent variables were scaled according to the criterion-based method with squared factor loadings using the criterion number of one (Schweizer, Troche, & Reiß, 2017).

2.3 Results

The mean fit results observed in investigating the generated data by means of the one-factor and two-factor models are included in Table 1.

Table 1. Mean Fit Results Obtained by One-factor and Two-factor Confirmatory Factor Models for Increasing Ranges of Difficulties of Items (N = 200)

R[1]	χ^2	RMSEA	SRMR	CFI	TLI	AIC	$d\chi^2$ (2 F)	d_{CFI} (2 F)	$d\chi^2$ (2 cF)	d_{CFI} (2 cF)
.5-.5	189.5	0.006	0.042	0.99	1.00	231.5	-	-	-	-
.4-.6	190.0	0.006	0.042	0.99	1.00	232.0	1.02	0.00	1.87	0.00
.3-.7	188.5	0.005	0.042	0.99	1.00	230.5	1.42	0.00	1.93	0.00
.2-.8	188.3	0.005	0.042	0.99	1.00	227.6	1.17	0.00	2.51	0.00
.1-.9	191.1	0.007	0.043	0.98	0.99	233.1	1.25	0.00	2.85	0.00

[1] Lower and upper limits of the range for difficulties of items.

The results reported in the columns two to seven were obtained by the one-factor model, in the columns eight and nine by the two-factor model with no correlation among the factors, and in the column ten and eleven by the two-factor model allowing for a correlation among the factors. All statistics for the one-factor model signified good model fit. The comparison between the CFIs of the one-factor models did not yield a substantial result (Note. Before rounding the statistics, the difference between CFI(.5-.5) and CFI(.1-.9) was smaller than .01). Furthermore, all chi-square and CFI differences reported in the columns eight to eleven indicated that there was no improvement in model fit.

The scaled mean variances are presented in Table 2.

Table 2. Mean Scaled Variances of Latent Variables of the One-factor [F1] and Two-factor [F1,F2] Confirmatory Factor Models for Different Ranges of Difficulties (N = 200)

R[1]	Var(F1) [F1]	Var(F1) [F1,F2]	Var(F2) [F1,F2]	Sum[2]	Var(F1) [cF1,cF2]	Var(F2) [cF1,cF2]	Cov [cF1,cF2]	Sum[3]
.5-.5	0.51	-	-	-	-	-	-	-
.4-.6	0.50	0.48	0.02	0.50	0.58	0.12	-0.10	0.50
.3-.7	0.46	0.45	0.02	0.47	0.51	0.07	-0.06	0.46
.2-.8	0.40	0.39	0.02	0.41	0.46	0.07	-0.06	0.41
.1-.9	0.33	0.31	0.02	0.33	0.37	0.06	-0.05	0.33

[1] Lower and upper limits of the range for difficulties of items.
[2] The sum (=whole variance at latent level) is the sum of the two variances.
[3] The sum (=whole variance at latent level) is the sum of the two variances plus the covariance (see Equation 11).

The second column includes the variances for the one factor model, the columns three and four for the two-factor model with no correlation among the factors, and the columns six and seven for the two-factor model allowing for a correlation among the factors. The covariance is reported in column eight. Furthermore, there are the columns five and nine that comprise the whole variances at the latent level.

The variances for the one-factor model revealed that there was a gradual decrease from 0.51 to 0.33 as a consequence of the increase in the range of the difficulty of the items. Furthermore, the comparison of these variances with the whole variances for the two-factor models showed that there was virtually constancy. There was no deviation larger than 0.01. Moreover, it became obvious that allowing the factors to correlate among each other led to a considerable increase of the variance of the first factors that was compensated by a negative covariance.

2.4 Discussion

The results of the simulation study did not support the expectation that increasing the range of difficulties would impair model fit and provide the basis for another factor that showed the characteristics of a difficulty factor. Instead all fit results suggested that there was one basic source of responding, as was induced in the process of generating the data.

However, the increase of the range of difficulties was not without an effect. The increase influenced the results in an unexpected way. There was a decrease in the size of the variance estimate for the factor of the one-factor model. This variance at the latent level showed to be virtually constant irrespective of the number of considered factors and the correlation among them.

3. The Effect of the Same High and Low Degree of Difficulty of the Items – Study 2

3.1 Introduction

The second study focused on the similarity of the items according to difficulty since a high degree of similarity was also proposed as the source of the difficulty factor (Bandalos & Gerster, 2016). Such similarity could be achieved by assigning the same degree of difficulty to several items. Since assigning the same degree to all items would mean uniformity that is assumed for the genuine factor, it appeared to be necessary to have two subsets of items that differed according to difficulty. But the difficulties of the items of each subset were the same.

3.2 Method

The methods of the first and second study only differed according to the cutoffs used in dichotomizing the continuous and normally distributed data. The 20 columns of the generated 500 × 20 matrices were subdivided into two subsets of ten columns. The columns of each subset were dichotomized according to the same cutoff. Four pairs of cutoff were employed for dichotomizing the data. The first one yielded difficulties of simulated items of 0.4 and 0.6, the second one of 0.3 and 0.7, the third one of 0.2 and 0.8, and the fourth one of 0.1 and 0.9. Furthermore, there were the matrices with columns dichotomized to yield simulated difficulties of 0.5. As in Study 1, the results observed for these matrices served as comparison level.

For further information see the method section of Study 1.

3.3 Results

Table 3 includes the mean fit results observed in investigating the generated data by means of the one-factor and two-factor models.

Table 3. Mean Fit Results Obtained by One-factor and Two-factor Confirmatory Factor Models for Increasing Distances Between Subsets of Items Showing the Same Difficulty (N = 200)

L[1]	χ^2	RMSEA	SRMR	CFI	TLI	AIC	$d\chi^2$ (2 F)	d_{CFI} (2 F)	$d\chi^2$ (2 cF)	d_{CFI} (2 cF)
.5/.5	189.5	0.006	0.042	0.99	1.00	231.5	-	-	-	-
.4/.6	187.9	0.006	0.042	0.99	1.00	230.9	0.94	0.00	1.83	0.00
.3/.7	186.3	0.005	0.042	0.99	1.00	228.3	1.35	0.00	2.75	0.00
.2/.8	185.6	0.005	0.042	0.99	1.00	227.6	1.08	0.00	3.70	0.00
.1/.9	192.6	0.007	0.043	0.96	0.99	234.6	2.92	0.00	7.71	0.01

[1] Lower and upper difficulties characterizing the subsets of items.

This Table shows the same structure as Table 1. The results reported in the first row correspond to the results reported in the first row of Table 1. As is obvious from the results reported in columns three to six, the one-factor model yielded good model fit in all cases. However, the comparison of the CFIs included in column 5 revealed a substantially worse model fit for .1/.9 as the pair of difficulty levels than for any other pair and the comparison level. Furthermore, only for this pair of difficulties the replacement of the one-factor model by the two-factor model led to improvement in model fit according to CFI and also to χ^2.

Table 4 comprises the scaled mean variances.

Table 4. Mean Scaled Variances of the Latent Variables of the One-factor [F1] and Two-factor [F1,F2] Confirmatory Factor Models for Different Subsets of Difficulties of Items (N = 200)

L[1]	Var(F1) [F1]	Var(F1) [F1,F2]	Var(F2) [F1,F2]	Sum[2]	Var(F1) [cF1,cF2]	Var(F2) [cF1,cF2]	Cov [cF1,cF2]	Sum[3]
.5/.5	0.51	-	-	-	-	-	-	-
.4/.6	0.48	0.49	-0.01	0.48	0.59	0.09	-0.10	0.48
.3/.7	0.39	0.36	0.02	0.38	0.44	0.07	-0.06	0.39
.2/.8	0.25	0.23	0.02	0.25	0.29	0.06	-0.05	0.25
.1/.9	0.10	0.09	0.01	0.10	0.12	0.03	-0.02	0.11

[1] Lower and upper difficulties characterizing the subsets of items.

[2] The sum (=whole variance at latent level) is the sum of the two variances.

[3] The sum (=whole variance at latent level) is the sum of the two variances plus the covariance (see Equation 11).

This Table shows the same structure as Table 2. Again a gradual decrease of the variance estimates for the one-factor model was observed as a consequence of the increase in the range of the difficulty of the items. It was an even stronger decrease than the decrease reported in Study 1 since the decrease from 0.51 ended with 0.10. Again the whole variances at the latent level virtually corresponded to the variance estimates for the one-factor model.

3.4 Discussion

This Study revealed one condition giving rise to results that signified the observation of a difficulty factor. There was the impairment in model fit and the evidence in favor of an additional factor that showed factor loading related to the difficulties of the items. The special features of this condition were equal sizes of the difficulties of the items of each subset of items. Equal sizes meant a high degree of similarity of the difficulties of the items, as was suggested to drive a difficulty factor by Bandalos and Gerster (2016).

The validity of this finding was impaired by the fact that impairment in model fit was signified by CFI (Bentler, 1990) in data due to items showing extremely high difficulties. The variances of binary items with extremely high difficulties were very small. As a consequence, the implicit comparison with the independence model was unlikely to reveal a major difference that would signify good model fit (Schweizer, Troche, & Reiß, 2019). Therefore, it was important that also the comparison of the one-factor and two-factor models by the chi-square difference additionally confirmed the observation of a difficulty factor.

4. The Effect due to Items Showing the Same High Difficulty - Study 3

4.1 Introduction

This study served the realization of the third objective that was inspired by the outcome of Study 2. The effect due to the combination of equal difficulties and of very large sizes of difficulties of items was to be investigated in more detail. Since the high difficulties of items were more important for the characterization of an additional factor than the small difficulties, the subset of items showing high difficulties was retained from Study 2. It remained to be investigated which characteristics of the other items would contribute to the observation of a difficulty factor. Therefore, we manipulated the difficulties of the other items.

4.2 Method

The methods employed as part of this study corresponded to the methods selected for the studies 1 and 2 with the exception of the dichotomization of the continuous and normally distributed data. The 20 columns of the generated 500 × 20 matrices were subdivided into two subsets of ten columns. The columns of each subset were dichotomized according to the same cutoff. The columns of one subset were dichotomized by a cutoff that yielded difficulties of items of 0.9. The columns of the second subset were dichotomized according to different cutoffs to achieve different levels. In the first level all columns of the second subset were dichotomized to achieve a difficulty of 0.8 (.8/.9), in the second level of 0.7 (.7/.9), in the third level of 0.6 (.6/.9), in the fourth level of 0.5 (.5/.9), in the fifth level of 0.4 (.4/.9), in the sixth level of 0.3 (.3/.9), in the sevens level of 0.2 (.2/.9) and in the eights level of 0.1 (.1/.9). Furthermore, the matrices with all columns dichotomized to yield difficulties of 0.5 were considered as comparison level.

Further information on methods can be found in the method section of Study 1.

4.3 Results

The investigation of the data by the one-factor and two-factor models led to the mean fit results reported in Table 5.

Table 5. Mean Fit Results Obtained by One-factor and Two-factor Confirmatory Factor Models for a Subsets of Items Showing the Same High Difficulty and Another Subsets of Items with Smaller Difficulties (N = 200)

L[1]	χ^2	RMSEA	SRMR	CFI	TLI	AIC	$d\chi^2$ (2 F)	d_{CFI} (2 F)	$d\chi^2$ (2 cF)	d_{CFI} (2 cF)
.5/.5	189.5	0.006	0.042	0.99	1.00	231.5	-	-	-	-
.8/.9	237.1	0.022	0.048	0.90	0.90	279.1	3.09	0.00	4.10	0.00
.7/.9	225.3	0.019	0.047	0.93	0.93	268.3	2.81	0.00	4.49	0.00
.6/.9	219.8	0.016	0.046	0.94	0.94	261.8	7.80	0.01	10.41	0.02
.5/.9	207.5	0.013	0.045	0.96	0.96	249.5	6.61	0.01	12.14	0.01
.4/.9	204.4	0.011	0.045	0.96	0.97	246.4	5.12	0.01	12.02	0.02
.3/.9	199.7	0.010	0.044	0.97	0.97	241.7	2.68	0.00	11.02	0.01
.2/.9	194.0	0.008	0.043	0.97	0.98	236.0	1.33	0.00	9.03	0.01
.1/.9	192.6	0.007	0.042	0.96	0.99	234.6	2.92	0.00	7.71	0.01

[1] Lower and upper difficulty levels characterizing the subsets of items.

The results reported in this Table are arranged in the same way as the results of Table 1. The first row includes the results for the comparison level that were also included in the first row of the Tables 1 and 3. The following rows provide the new results with the exception of the last row. This row includes results already reported in the last row of Table 3.

Virtually all results for the one-factor model reported in the columns three to six suggested impairment of model fit due to the manipulation of the difficulties of items. All CFI differences for comparisons with the comparison level (.5/.5) were larger than 0.1. The comparison of the AICs yielded the smallest AIC for the comparison level. Furthermore, the smallest RMSEA was found for the model of the comparison level and also the smallest SRMR with one exception.

The consideration of an additional uncorrelated factor led to a substantial improvement in model fit in three levels (.6/.9, .5/.9, .4/.9) according to the chi-square difference and also the CFI difference. Allowing the two factors of the two-factor model to correlate with each other led to substantial chi-square differences for all levels. The CFI difference also reached the level of significance for comparisons of the CFI for the comparison level with the CFIs of six other levels (.6/.9, .5/.9, .4/.9, .3/.9, .2/.9, .1/.9).

The means estimates of the variance parameter are presented in Table 6 (see next page). The structure of this Table corresponds to the structure of Table 2. The comparison of the entries of the second column including the variance estimates for the one-factor model revealed that the estimate for the comparison level was almost two times as large as any other estimate. The comparison of the variance estimates reported in the third and sixth columns that referred to the genuine factor of the two-factor model showed to be considerably larger than the estimate for the corresponding one-factor model. In the .8/.9 and .7/.9 levels the estimate for the correlated genuine factor was more than four times as large as the estimate for the factor of the corresponding one-factor model. These large sizes were compensated by negative variance and covariance estimates since in all cases the whole variance at the latent level virtually corresponded to the variance estimate for the factor of the corresponding one-factor model.

4.4 Discussion

All levels considered in this study appeared to drive a difficulty factor. A common feature of these levels was the equality of the difficulties and also the very large difficulties of one subset of items. The likelihood of observing a difficulty factor appeared to be especially high if the other items showed difficulties near to the mean of the range between zero and one. Apparently complex patterns of difficulties of items led to the observation of

Table 6. Mean Scaled Variances of Latent Variables of the One-factor [F1] and Two-factor [F1,F2] Confirmatory Factor Models for the Subset with Constant Diffculties and Other Subsets of Difficulties (N = 200)

D[1]	Var(F1) [F1]	Var(F1) [F1,F2]	Var(F2) [F1,F2]	Sum[2]	Var(F1) [cF1,cF2]	Var(F2) [cF1,cF2]	Cov [cF1,cF2]	Sum[3]
.5/.5	0.51	-	-	-	-	-	-	-
.8/.9	0.18	0.39	-0.20	0.19	0.88	0.28	-0.49	0.18
.7/.9	0.23	0.37	-0.16	0.21	0.94	0.30	-0.50	0.24
.6/.9	0.26	0.39	-0.12	0.27	0.82	0.20	-0.37	0.27
.5/.9	0.26	0.33	-0.07	0.26	0.66	0.13	-0.26	0.27
.4/.9	0.25	0.29	-0.04	0.25	0.51	0.08	-0.16	0.27
.3/.9	0.22	0.23	-0.01	0.22	0.37	0.06	-0.11	0.23
.2/.9	0.16	0.16	-0.00	0.16	0.24	0.04	-0.06	0.16
.1/.9	0.10	0.09	0.01	0.10	0.12	0.03	-0.03	0.09

[1] Lower and upper difficulty levels characterizing the subsets of items.
[2] The sum (=whole variance at latent level) is the sum of the two variances.
[3] The sum (=whole variance at latent level) is the sum of the two variances plus the covariance (see Equation 11).

a difficulty factor. These patterns were to some degree in line with each one of the ideas regarding the source of the difficulty factor (Bandalos & Gerster, 2016; Ferguson, 1941; McDonald, 1965, and others). But no one of the ideas seemed to reflect the complexity completely.

The large increases of the estimates of the variance parameter for the genuine factor were unexpected results. In some cases the size of this estimate for the genuine factor of the two-factor model surmounted the size of the estimate for the corresponding factor of the one-factor model several times. This increase occurred although the latent source driving the genuine factor was kept constant. The large variances were compensated by either a negative estimate of the variance of the additional factor or a negative estimate of the covariance. Whereas the covariance could be negative, variances must be positive (Stoel, Garre, Dolan, & van de Wittenboer, 2006; Verbeke & Molenberghs, 2003). But there was also no reason for expecting a negative correlation among the factors that might lead to a negative covariance.

4. General Discussion

Factor-analytic models distinguish between latent and observed variables and assume that the latent variables influence the observed variables (Flora, LaBrish, & Chalmers, 2012). Furthermore, the latent variables that are also referred to as factors usually represent latent sources. The responses to the items that load on a factor are assumed to be influenced by the corresponding latent source. A factor that is not driven by such a source is an annoyance since the researchers' effort to disclose the meaning of such a factor is likely to fail. Such an experience led Guilford (1941) to provide the first report of a difficulty factor and to open up a new field of research this way.

For avoiding such situations, researchers need to be equipped with methods for identifying factors that are not driven by a latent source as, for example, the difficulty factor. One method proposed for the identification of a difficulty factor is the investigation of the relationship between the difficulties of the items and the sizes of the factor loadings since a difficulty factor is known to show factor loadings according to the difficulties of the items. However, as pointed out by McDonald and Ahlawat (1974), such a relationship does not exclude the possibility that there is an underlying source driving the additional factor. For example, the item-position effect is related to the difficulties of the items if the items are arranged according to their difficulties (Schweizer Troche, 2018). But the corresponding factor is a factor that is independent of the difficulties (Zeller, Reiss, & Schweizer, 2017).

The reported research work aims at the identification of the conditions that lead to the observation of a difficulty factor. It relies on simulated data since such data enable the systematic manipulation of the properties characterizing the data. Furthermore, the investigation makes use of fixed-links models (Schweizer, 2006, 2008, 2009) that enable the theory-guided decomposition of the variances and covariances. The factor loadings of such a model can be specified according to the difficulties of the items in order to assure that the corresponding factor shows a basic characteristic of a difficulty factor: a close relationship of the difficulties of the items and the sizes of the factor loadings.

The results of the research work add to the knowledge on the conditions that drive a difficulty factor. The first study reveals that the broad range of difficulties of the items is not as important, as is suggested in literature (Ferguson, 1941; McDonald & Ahlawat, 1974). The second study signifies that very high and low difficulties of equal size may drive the difficulty factor (Bandalos & Gerster, 2016). The third study shows what actually leads to the difficulty factor: there must be one subgroup of items showing the same high difficulty. This means inhomogeneity with respect to the complete set of items (Bandalos & Gerster, 2016). Each one of the levels of this study includes a subgroup of items with very high difficulties of equal size whereas the difficulties of the other items vary.

The knowledge about the condition driving a difficulty factor cannot only be used for the identification of a difficulty factor but also for avoiding such a factor. Guidelines for item writing and item selection (Johnson & Morgan, 2016) should equip researchers with knowledge appropriate for this purpose. In real data the effect of equal difficulties may vary as a function of the type and number of processes contributing to performance. The effect may be stronger in simple items than in complex items since fewer processes contribute. In contrast, in complex items the effect may be weak since more processes that may differ from each other are likely to contribute (Schweizer, 1998, Stankov, 2001).

Furthermore, there is the additional information provided by the variance estimates for the latent variables. Variances of latent variables are rarely considered in empirical research with the exception of longitudinal research and invariance analyses (McArdle, 2009; McArdle & Cattell, 1994; Schmitt & Kuljanin, 2008). Variances can be useful if the model of measurement includes more than one latent variable since in this case a comparison of the contributions of the latent variables to the explanation of the criterion variable is possible. In the present case the comparisons of the variances for the latent variable of the one-factor model and the whole variance of the two-factor models reveals virtual equality. This result is in line with the presumption that there is no additional underlying source for the difficulty factor.

Although the research work provides new insight regarding the conditions driving the difficulty factor and the consequences for the variance associated with the latent variables, there is also a limitation. The limitation is the lack of a contribution to the explanation of how the difficulty factor is generated. Further research is necessary to improve our understanding of the mechanism that leads to the difficulty factor.

Acknowledgments

There was no funding.

References

Bandalos, D. L., & Gagné, P. (2012). Simulation methods in structural equation modeling. In R. H. Rick (Ed.), *Handbook of structural equation modeling* (pp. 92-108). New York, NY: Guilford Press.

Bandalos, D. L., & Gerstner, J. J. (2016). Using factor analysis in test construction. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction* (pp. 26-51). Göttingen, Germany: Hogrefe.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246. https://doi.org/10.1037/0033-2909.107.2.238

Byrne, B. M. (2016). Using multitrait-multimethod analysis in testing for evidence of construct validity. In K. Schweizer & C. Distefano (Eds.), *Principles and Methods of Test Construction* (pp. 288-307). Göttingen: Hogrefe Publishing.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255. https://doi.org/10.1207/S15328007SEM0902_5

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: a systematic review of issues and developments. *Front. Psychol, 9,* 1-14. https://doi.org/10.3389/fpsyg.2018.00580

DiStefano, C. (2016). Examining fit with structural equation models. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction* (pp. 166-193). Göttingen: Hogrefe.

Ferguson, G. A. (1941). The factorial interpretation of test difficulty. *Psychometrika, 6,* 323-329. https://doi.org/10.1007/BF02288588

Flora, D. B., LaBrish, C., & Chalmers, R. P. (2012). Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. *Frontiers in Psychology* (Section Quantitative Psychology and Measurement). https://doi.org/10.3389/fpsyg.2012.00055

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286-299. https://doi.org/10.1037/1040-3590.7.3.286

Gibson, W. A. (1959). Three multivariate models: factor analysis, latent structure analysis, and latent profil analysis. *Psychometrika*, *24*, 229-252. https://doi.org/10.1007/BF02289845

Gibson, W. A. (1960). Nonlinear factors in two dimensions. *Psychometrika, 25*, 381-392. https://doi.org/10.1007/BF02289755

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement, 66*, 930-944. https://doi.org/10.1177/0013164406288165

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika, 6*, 67-77. https://doi.org/10.1007/BF02292175

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. https://doi.org/10.1080/10705519909540118

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika, 36*, 109-133. https://doi.org/10.1007/BF02291393

Jöreskog, K. G., & Sörbom, D. (1999). *PRELIS 2: User's reference guide*. Lincolnwood, IL: Scientific Software International Inc.

Jöreskog, K. G., & Sörbom, D. (2001). *Interactive LISREL: User's guide*. Lincolnwood, IL: Scientific Software International Inc.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80*. Lincolnwood, IL: Scientific Software International Inc.

Johnson, R. L., & Morgan, G. B. (2016). Item types, response formats, and consequences for statistical investigations. In K. Schweizer & C. DiStefano (Eds.), *Principles and methods of test construction* (pp. 83-103). Göttingen: Hogrefe.

Kishton, J. M., & Widaman, K. F. (1994). Unidimensional versus domain representative parceling of questionnaire items: An empirical example. *Educational and Psychological Measurement, 54,* 757-765. https://doi.org/10.1177/0013164494054003022

McArdle, J. J. (2009). Latent variable modeling of differences and change with longitudinal data. *Annual Review of Psychology, 60*, 577-605. https://doi.org/10.1146/annurev.psych.60.110707.163612

McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research, 29,* 63-113. https://doi.org/10.1207/s15327906mbr2901_3

McCullagh, P., & Nelder, J. A. (1985). *Generalized linear models*. London: Chapman and Hall.

McDonald, R. P. (1965). Difficulty factors and nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology, 18*, 11-23. https://doi.org/10.1111/j.2044-8317.1965.tb00690.x

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27*, 82-99. https://doi.org/10.1111/j.2044-8317.1974.tb00530.x

Muthen, B. (1984). A general structural equation model with dichotomous, ordered, categorical, and continuous latent variables indicators. *Psychometrika, 32*, 1-13. https://doi.org/10.1007/BF02294210

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resources Management*, *18*, 210-222. https://doi.org/10.1016/j.hrmr.2008.03.003

Schweizer, K. (1998). Complexity of information processing and the speed-ability relationship. *The Journal of General Psychology, 125*, 89-102. https://doi.org/10.1080/00221309809595578

Schweizer, K. (2006). The fixed-links model for investigating the effects of general and specific processes on intelligence. *Methodology, 2*, 149-160. https://doi.org/10.1027/1614-2241.2.4.149

Schweizer, K. (2008). Investigating experimental effects within the framework of structural equation modeling: an example with effects on both error scores and reaction times. *Structural Equation Modeling*, *15*, 327-345. https://doi.org/10.1080/10705510801922621

Schweizer, K. (2009). Fixed-links models for investigating experimental effects combined with processing strategies in repeated measures designs: a cognitive task as example. *British Journal of Mathematical and Statistical Psychology, 62*, 217-232. https://doi.org/10.1348/000711007X268558

Schweizer, K. (2013). A threshold-free approach to the study of the structure of binary data. *International Journal of Statistics and Probability, 2*, 67-75. https://doi.org/10.5539/ijsp.v2n2p67

Schweizer, K., Ren, X., & Wang, T. (2015). A comparison of confirmatory factor analysis of binary data on the basis of tetrachoric correlations and of probability-based covariances: a simulation study. In R. E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research* (pp. 273-292). Heidelberg: Springer.

Schweizer, K., Ren, X., Wang, T., & Zeller, F. (2015). Does the constraint of factor loadings impair model fit and accuracy in parameter estimation? *International Journal of Statistics and Probability, 4*, 40-50. https://doi.org/10.5539/ijsp.v4n4p40

Schweizer, K., & Troche, S. (2018). Is the factor observed in investigations of the item-position effect actually the difficulty factor? *Educational and Psychological Measurement*, *78*, 46-69. https://doi.org/10.1177/0013164416670711

Schweizer, K., Troche, S., & DiStefano, C. (2019). Scaling the variance of a latent variable while assuring constancy of the model. *Frontiers in Psychology* (Section Quantitative Psychology and Measurement), *10*, ArtID 887. https://doi.org/10.3389/fpsyg.2019.00887

Schweizer, K., Troche, S., & Reiß, S. (2017). Can variances of latent variables be scaled in such a way that they correspond to eigenvalues? *International Journal of Statistics and Probability, 6*, 35-42. https://doi.org/10.5539/ijsp.v6n6p35

Schweizer, K., Troche, S., & Reiß, S. (2019). Does the effect of a time limit for testing impair structural investigations by means of confirmatory factor models? *Educational and Psychological Measurement*, *79*, 40-64. https://doi.org/10.1177/0013164418770824

Stankov, L. (2001). Complexity, metacognition and fluid intelligence. *Intelligence*, *28*, 121-143. https://doi.org/10.1016/S0160-2896(99)00033-1

Stoel, R. D., Garre, F. G., Dolan, C., & van de Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*, 439-455. https://doi.org/10.1037/1082-989X.11.4.439

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics, 59*, 254-262. https://doi.org/10.1111/1541-0420.00032

Zeller, F., Reiss, S., & Schweizer, K. (2017). Is the item-position effect in achievement measures induced by increasing item Difficulty? *Structural Equation Modelling, 24*, 745-754. https://doi.org/10.1080/ 10705511.2017.1306706