Psychological Test and Assessment Modeling, Volume 63, 2021 (4), 469-483

Conditions leading to the observation of a difficulty effect and its consequence for confirmatory factor analysis

Karl Schweizer[1], Christine DiStefano[2], and Stefan Troche[3]

Abstract

Psychometric research has posited that difficulty effects related to item-level input for factor analysis serves as the precondition for observing a difficulty factor. Two studies are reported that investigated and confirmed this hypothesis. First, it was demonstrated that extreme and same-sized extreme difficulty levels of items resulted in deviations of the input to factor analysis from the expected systematic variation. Difficulty levels, as defined by McDonald and Ahlawat (1974), that were close to the upper limit for such levels were used for this purpose. Subsequently, it was demonstrated that data with this effect were likely to show model misfit in structural investigations by the one-factor CFA model. According to these results the difficulty effect is a method effect caused by the difficult factor condition of data. This condition is the source of additional systematic variation that is not accounted for by the intended latent variable. This additional variation means model misfit unless it is captured by a difficulty factor.

Keywords: difficulty effect, difficulty factor, factor analysis, method effect

[1] Correspondence concerning this article should be addressed to Karl Schweizer, Institute of Psychology, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 6, 60323 Frankfurt a. M., Germany. E-mail: K.Schweizer@psych.uni-frankfurt.de

[2] Department of Educational Studies, University of South Carolina

[3] Department of Psychology, University of Bern

In psychometric research, the label difficulty factor is used to characterize factors where loading values reflect the difficulty levels of the items, but the values do not provide a clear picture of the content measured by the factor (Guilford, 1941). Since such a factor seems to serve no obvious purpose, it is usually considered as a reflection of construct irrelevant variance that does not contribute to the understanding of the content. Researchers have debated a number of possible circumstances that may give rise to such a factor, such as a broad range of difficulty values, non-linear underlying dimensions or ways of computing correlations (Bandalos & Gerstner, 2016; Ferguson, 1941; Gibson, 1959, 1960; McDonald, 1965, 1967; McDonald & Ahlawat, 1974; Wherry & Gaylord, 1944). The research reported in this paper differs from the previous approaches in that it investigates the possibility of a difficulty effect as a specific method effect (for details see next paragraph) that may give rise to a difficulty factor. This approach proceeds from the suggestion that extreme difficulty levels of items may play a key role in the generation of a difficulty factor (Bandalos & Gerstner, 2016). The effect of extreme difficulty levels of items on the input to factor analysis and also on the results of factor analysis is investigated in simulated data.

Data characteristics resulting from the measurement process, which are unrelated to what an instrument is assumed to measure, are considered as method effects (Maul, 2013; Schweizer, 2020). These effects are by-products of the stimulation of a specific response in measurement so that a specific source can be expected as the origin of the response. Such effects are apparent whenever the outcome of measurement not only reflects the intended source but an influence of an additional source of variation. Accordingly, Sechrest, Davis, Stickle and McKnight (2000, p. 64) describe a method effect as an effect which does not originate from the attribute to be measured but shows to be related to the observational procedure. This argument can be extended to the context of measurement because the context can influence measurement in a characteristic way. For example, the method effect referred to as speededness (Oshima, 1994) is the consequence of an insufficient time span for completing the items of an achievement scale, resulting in omitted items. In the case of speededness, some participants do not reach their optimal score, due to a time limit imposed on the testing situation. In sum, method effects bias observed data and, thereby, impair the quality of measurement (Lu & Sireci, 2007).

The structure of a scale is most frequently examined by means of factor analysis. For this purpose, the outcome of measurement is transformed into the input to factor analysis (e.g., an empirical variance-covariance matrix). Consequently, method effects included in the input to factor analysis may influence the outcome of factor analysis. For example, factor analysis expects complete data, but speededness may have caused omissions. An accepted way of dealing with omissions is data imputation (Graham, 2009). But, as there is no guarantee that the replacement of an omission exactly corresponds to what is missing, the outcome of factor analysis may still show an influence of the method effect. Other method effects may demand other provisions.

What is expected as input to factor analysis is complete structured random data showing interval scale and a specific distribution (i.e., an approximately normal distribution, Graham 2006). But even if data meet the mentioned expectations, they may still be influenced by a method effect because they may include systematic variation other than the systematic variation originating from the stimulated source of responding. Unlike random influences, method effects are sources of additional systematic variation of data. For example, there is the item-position effect (Knowles, 1988; Zeller, Reiß, & Schweizer, 2017). This effect is apparent in the dependency of item statistics on the positions of the items within the series of items, in the gradual increase of the item reliability and the correlations among the items. It is also apparent as additional systematic variation of items that increase along the sequence of the arrangement of items. It is such an additional kind of variation that is captured by a difficulty factor in factor analyses.

Taken together, this discussion suggests that a proper understanding of the difficulty factor requires the identification of the source of additional systematic variation, the investigation of the type of additional systematic variation and the consequences for factor analysis. Knowledge of what leads to a difficulty effect means that it is possible to predict the observation of a difficulty factor.

Study 1: The Difficulty Effect

In this study we investigate the characteristics of data that, when used as input for a factor analysis, are likely to give rise to a difficulty factor. Since in the following discussion the concept of the difficulty level that is also addressed as level of difficulty plays a major role, we like to clarify the meaning of this concept before going into details. We follow McDonald and Ahlawat (1974) who define difficulty level as level "… measured by the proportion of examinees passing each item" (p. 84). This means that a high difficulty level of an item signifies easiness. Such an item is an easy item since many participants are able to complete it correctly. Furthermore, it needs to be pointed out that very difficult items are likely to show the same characteristic as very easy items in that both these types of items tend to show rather small variances. As variances and covariances or correlations serve as input for factor analysis, both items with high and low difficulty levels may give rise to a kind of difficulty factor. Therefore, it may be more appropriate to refer to extreme items as source of such a factor instead of very difficult items only.

In early research on difficulty factors, many researchers favored the hypothesis that variability due to a wide range of item difficulty levels led to the presence of a difficulty factor (Ferguson, 1941; Gibson, 1959, 1960). Although this hypothesis did not find general acceptance, there has been agreement that variability of item difficulty values may be of importance as a precondition of a difficulty factor. More recent literature suggests that within a set of items, items with extreme difficulty levels and especially items that have the same extreme level of difficulty will produce a difficulty factor (Bandalos & Gerstner, 2016). A simulation study with datasets that comprised a few items with extreme difficulty levels and also same-sized extreme difficulty levels among other items demonstrated that the presence of such items increased the number of observed difficulty factors (Schweizer & Troche, 2018). In the following we refer to this specific set of characteristics as *difficulty factor condition*.

Another characteristic likely to give rise to a difficulty factor relates to the scale of the items, where only binary data seem to result in a difficulty factor (Floyd & Widaman, 1995; McDonald & Ahlawat, 1974). With binary data, the variance of an item and the probability of a correct response are closely related. Furthermore, it is known that the loadings on a difficulty factor reflect the difficulty levels of the items that are closely linked to the probabilities of solving these items (Guilford, 1941). If the number of ordered categories is increased, this linkage is weakened or even disappears. As a consequence, when based on binary data, factor loadings are highly correlated with the variances used as (part of the) input to factor analysis.

Although the aforementioned conditions may contribute to observing a difficulty factor, they do not explain how this factor may arise. Such an explanation requires the investigation of how the difficulty factor condition influences the input to factor analysis. The input is usually a correlation or covariance matrix. We restrict the investigation to correlations because of the better interpretability.

Objectives

Taking the described approach of explaining a difficulty factor, the question to be investigated can be specified as follows: what are the special characteristics of the correlations computed from pairs of items with extreme and equal-sized extreme difficulty levels if the data are binary?

Method

To elucidate this situation, we generated 300 datasets of binary and structured random data with sample sizes of 500. Each dataset included five items (= columns) generated such that correlations between items of .1225 were expected. This means that in factor analysis these correlations would be reproducible by factor loadings of 0.35. Two items (= columns) showed difficulty levels of .75, another one of .80 and the remaining two items of .95. We refer to the items with difficulty levels of .75 and .80 as *normal difficult* and the other items as *extreme difficult*. The difficulty levels of .75 and .80 were taken from the difficulty level range between .20 and .80 that was found to have no effect on the probability of observing a difficulty factor (Schweizer & Reiß, 2019).

Prelis (Jöreskog & Sörbom, 2001) was used to generate data and to compute tetrachoric correlations that were recommended for estimating relationships of binary data. The estimated correlations were subdivided according to size, and assigned into one of eight correlational ranges ([-1 to -.75[, [-.75 to -.5[, [-.5 to -.25[, [-.25 to 0[, [0 to .25],].25 to .5],].5 to .75],].75 to 1]). Note. In order to clearly define to which interval a boundary value (e.g., .25) has to be assigned, we use brackets for inclusion and exclusion: in the case of […] the interval includes the boundary values; in the case of]…[the interval only includes estimates that approximate the boundary values; and there are combinations of both types ([..[and]..]) for combining inclusion and exclusion. Subsequently, the estimates assigned to the ranges (= groups) were counted. It was expected that the estimates would fall exclusively into the range between 0 and .25, as .1225 was the expected correlation.

Results

Figure 1 illustrates the results.

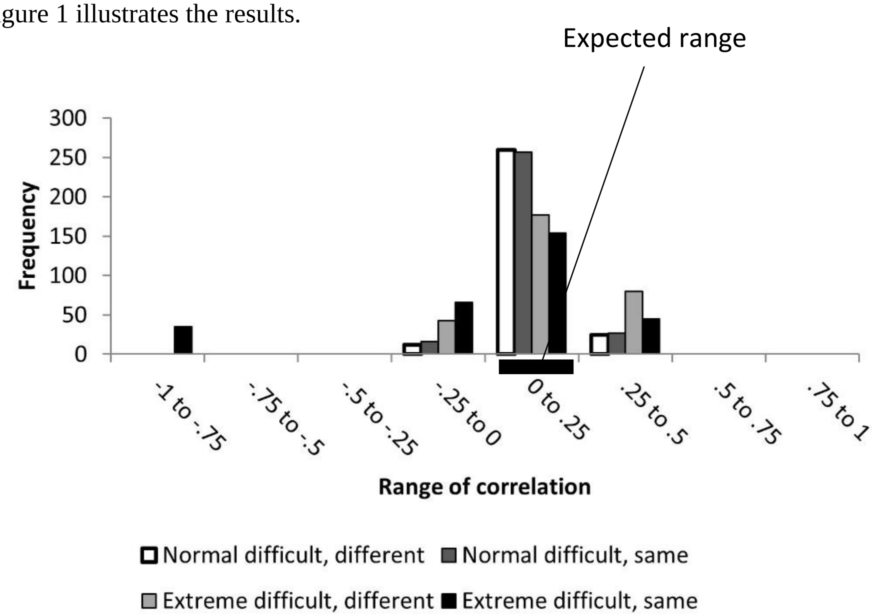

Figure 1. Frequencies of correlations in the expected range (0 to .25) and other ranges for pairs of items showing different difficulty levels (normal difficult, different: .75 and .80; normal difficult, same: both .75; extreme difficult, different (= wide difficulty range): .75 and .95; extreme difficult, same: both .95). Each level included the computation and investigation of 300 tetrachoric correlations.

Across the set of simulated difficulty conditions, the normal difficulty data conditions [pairs of items denoted as *normal difficult, different* (levels of .75 and .80) and *normal difficult, same* (levels of .75 and .75)] showed the best agreement between estimated and expected values. For these conditions, the expected range included roughly 86 percent of the tetrachoric correlations. In contrast, only 59 percent of the observations for items under a wide range of difficulty levels (levels of .75 and .95) fell in the expected range, and only 51 percent of the observations for items with extreme and same-sized extreme difficulty levels (levels of .95 and .95). For pairs of items of same-sized extreme difficulty level, some very large negative correlation values were observed that might be due to problems in threshold estimation as part of the computation of tetrachoric correlations. Moreover, the results for items with same-sized extreme difficulty levels and for items with a wide range of difficulty levels differed. With same-sized extreme values, the majority of the deviations (between observed and expected values) were negative, and with a wide range of difficulty values, deviations were positive.

Furthermore, we checked whether the frequency distributions for the different types of items corresponded or differed substantially from each other. We combined the normal difficult, different and normal difficult, same observations on the one hand and the extreme difficult, different and extreme difficult, same observations on the other hand for conducting a χ^2 test. This test was restricted to the three ranges with a larger number of observations ([-.25 to 0[, [0 to .25],].25 to .5]). A χ^2 of 403.1 (df = 2) was observed that suggested different distributions.

Discussion

In sum, extreme difficulty levels resulted in larger negative and positive deviations from the expected size of the tetrachoric correlation than moderate difficulty levels. This meant that these observed values would not be reproduced well by a factor with factor loadings that could be expected to reproduce correlations falling into the expected range. Equal-sized difficulty levels even increased the probability of a large deviation value (especially in the negative direction). The effect of equal-sized difficulty levels was probably due to the increased chance of perfect correspondence of the subsets of correct and incorrect responses of two items. Overall, the results of investigating the effect of the difficulty factor condition on the input to factor analysis revealed changes. These were changes leading away from characteristics of data that could be expected to lead to good model fit.

The results show that the difficulty factor condition produces deviations from expected sizes of correlations (or covariances). This means that a modification of an otherwise homogeneous set of correlations (or covariances) is created. This modification is, however, restricted to the subset of items showing the difficulty factor condition. This suggests that the *difficulty effect* is a systematic deviation of some coefficients (correlations or covariances) from an otherwise rather homogeneous pattern of correlations or covariances that is associated with the intended source of responding. The described systematic deviation due to the difficulty factor condition can alternatively be perceived and addressed as additional systematic variation that is restricted to a subset of items showing the difficulty factor condition.

Study 2: Influence of the Difficulty Effect on Confirmatory Factor Analysis

This study addresses the consequences of the difficulty factor condition for customary confirmatory factor analysis using the one-factor confirmatory factor analysis model. At first, we introduce the framework of the investigation: the confirmatory factor analysis (CFA) model, the corresponding covariance matrix (CV) model and the scaling method for the variance parameter. Afterwards, we report an investigation on how the difficulty factor condition influences model fit and the scaled variance of the latent variable when using a one-factor CFA model in combination with free and fixed factor loadings.

The one-factor model of measurement of customary CFA (Brown, 2015) is a linear combination of two components that are considered as a systematic component and an error component since they refer to the systematic variation of data on one hand and the error variation of data on the other hand. This one-factor CFA model is described by the following equation:

$$\mathbf{x} = \boldsymbol{\lambda}\xi + \boldsymbol{\delta} \qquad (1)$$

where the $p \times 1$ vector $\mathbf{x}$ represents the manifest variables, the $p \times 1$ vector $\boldsymbol{\lambda}$ the factor loadings, ξ the latent variable, and the $p \times 1$ vector $\boldsymbol{\delta}$ the error variables. This CFA model does not include a parameter that represents the variance of the latent variable. Such information on the latent variance, however, would be of interest because it could provide information on the systematic variation of data captured by the latent variable. The CV model is necessary for getting access to the latent variable variance (Jöreskog, 1970). The $p \times p$ model-implied covariance matrix, $\boldsymbol{\Sigma}$, for p manifest variables is defined as

$$\boldsymbol{\Sigma} = \boldsymbol{\lambda}\phi\boldsymbol{\lambda}' + \boldsymbol{\Theta} \quad . \qquad (2)$$

It includes the $p \times 1$ vector of factor loadings, $\boldsymbol{\lambda}$ (and its transpose $\boldsymbol{\lambda}'$), the variance parameter, ϕ, and the $p \times p$ diagonal matrix of error variation, $\boldsymbol{\Theta}$. Within this model the variance parameter, ϕ, represents the latent variable variance. It is a scalar. The two components of this model serve the representation of the systematic variation of data, on the one hand, and of random error, on the other hand (Schweizer, Troche, & DiStefano, 2019).

In data showing a unidimensional underlying structure the investigation of model fit by comparing the $p \times p$ empirical covariance matrix, $\mathbf{S}$, with the model-implied $p \times p$ covariance matrix, $\boldsymbol{\Sigma}$, can be expected to yield good model fit if the model is specified to expect one factor. In this case the systematic part of the model, $\boldsymbol{\lambda}\phi\boldsymbol{\lambda}'$, accounts for the systematic variation of data.

In contrast, in the case of the difficulty factor condition, signified by the subscript **DF**, the empirical matrix, $\mathbf{S}_{DF}$, is likely to show systematic variation according to two sources that are at least partly independent of each other. (1) There is systematic variation due to the main source of responding that finds its expression in the sizes of the correlations or covariances among all items of the correlation or covariance matrix that serves as input to factor analysis. (2) There is the subset of extreme items showing the difficulty factor condition which gives rise to additional variation that is apparent in the deviations of the correlations or covariances from what is expected according to the first source (see previous section).

Proceeding from the assumption that the data show a unidimensional structure, as it is the standard procedure for structural investigations, the empirical correlation or covariance matrix ($\mathbf{S}_{DF}$) is investigated by means of a one-factor CFA model that gives rise to a model-implied matrix according to Equation 2. This means that two-dimensional data are investigated by a model assuming one dimension. Therefore, we hypothesize that the discrepancy function F serving as ingredient of several fit indices signifies a larger deviation between $\mathbf{S}_{DF}$ and $\mathbf{\Sigma}$ as between $\mathbf{S}$ and Σ:

$$F[\mathbf{S}_{DF}, \mathbf{\Sigma}] > F[\mathbf{S}, \mathbf{\Sigma}] \quad . \qquad (3)$$

Furthermore, the fit indices used for evaluating the quality of model-data combinations are likely to signify impairment in model fit for the data showing the difficulty effect in comparison to the data without such an effect.

The factor loadings play the major role in capturing the systematic variation of data in exploratory factor analysis (Widaman, 2018). Although the outcome of factor analysis is mainly evaluated by inspecting the factor loadings, there is also the tradition to estimate the variance of a factor on the basis of factor loadings. The variance of the factor, var(ξ), combines what is captured by the individual factor loadings: the squared factor loadings are summed:

$$\mathrm{var}(\xi) = \sum_i \lambda_i^2 \quad . \qquad (4)$$

In contrast, the CV model of confirmatory factor analysis (Equation 2) includes the variance parameter, ϕ, that can represent the variance of the latent variable but also serve other purposes (e.g. serve as constant basis and reference point for estimating the variances of other latent variables) when scaled accordingly (Little, Slegers, & Card, 2006). Using it for representing systematic variation requires EV (eigenvalue[4]) scaling (Schweizer & Troche, 2019), where estimates of factor loadings achieved in the first analysis are transformed to meet the demands of EV scaling. In a new analysis with a free variance parameter, ϕ_{EV}, and fixed entries of the vector of factor loadings an estimate corresponding to var(ξ) is achievable:

$$\phi_{EV} = \mathrm{var}(\xi) \quad . \qquad (5)$$

If data are generated to show a unidimensional structure, the estimate of the scaled variance parameter reflects the systematic variation of $\mathbf{S}$ well (= $\mathbf{S}_{No\_DF}$). In contrast, if the data show the difficulty factor condition, there are positive or negative deviations from what is expected as systematic variation of data. This condition transforms

[4] This scaling method leads to estimates similar to eigenvalues if the manifest variables (e.g., items) show variances of one.

$\mathbf{S}_{No\_DF}$ into $\mathbf{S}_{DF}$ (see previous section). Because in the investigation of simulated data under the difficulty factor condition negative deviations show the overall stronger deviation from what is expected, we expect a decrease of the average of estimated factor loadings of the one-factor model. This means a decrease of the variance of the latent variable estimated by ϕ_{EV}:

$$\phi_{EV}\left(\mathbf{S}_{DF}\right) < \phi_{EV}\left(\mathbf{S}_{No\_DF}\right) \quad . \quad (6)$$

Objectives

This study aims at providing empirical evidence regarding the following hypotheses: (1) the difficulty factor condition leads to model misfit in the investigation of the structure of data that otherwise enable the observation of good model fit, as is suggested by Inequity 3. (2) Impairment in model fit due to the difficulty factor condition is accompanied by an impaired representation of the systematic variation of data, as is indicated by a decreased size of variance of the latent variable. This is expected because of deviations from the systematic variation due to the main latent source of responding especially in the negative direction. (3) Free and fixed factor loadings lead to the same degree of good model fit if the data show systematic variation due to the assumed latent source of responding only, as there is no additional systematic variation that can additionally be tapped by free factor loadings.

Method

In order to examine these hypotheses, we extended the investigation of the previous example by generating 300 datasets that were 500 × 20 data matrices instead of 500 × 5 matrices. Again data with an underlying structure giving rise to the expectation of factor loadings of 0.35 were generated (Jöreskog & Sörbom, 2001). In order to have two data types for a comparison, continuous, normally distributed data were transformed into binary data in two different ways. Input matrices were first transformed by allowing 17 items (= columns) to have difficulty levels of .75 and the remaining three items, of .95 (first way). The items with the high probability levels were assigned to the positions of the 5th, 10th and 15th columns. Otherwise, difficulty levels were set at .75 for all items (second way).

The data (i.e., matrices) were investigated by CFA models according to Equation 1, allowing for either free or fixed factor loadings. In the case of fixed factor loadings, the value of 0.223 (that follows from EV scaling: $1 = 20 \times 0.223^2$; this value is selected for the factor loadings that sum up to one after being squared) was assigned to all factor loadings to reflect the assumption that the latent source affects all manifest variables in the same way. To extend the previous empirical investigation, we used the tetrachoric correlation-based CFA approach (Muthén, 1984). This approach additionally required robust estimation in small datasets (Satorra & Bentler, 1994). The analyses were conducted using the LISREL software package (Jöreskog & Sörbom, 2006) with the ML option (maximum likelihood estimation).

Results

Table 1 includes average fit statistics and standard deviations resulting from investigating the simulated data by one-factor CFA models. The evaluation concentrated on RMSEA (root mean squared error of approximation), SRMR (standardized root mean residual), NNFI (nonnormed fit index) and CFI (comparative fit index) because of the availability of cutoffs (see DiStefano, 2016; Hu & Bentler, 1999). All RMSEA and SRMR statistics indicated good model fit as shown by values under suggested cutoffs of .05 and .08, respectively. The NNFI and CFI statistics also showed good model fit in the absence of the difficulty factor condition with values greater than a cutoff of .95; however, model fit was only acceptable (i.e., values close to .90) in the presence of this condition. In all cases, the difference between the CFIs for the presence and absence of the difficulty factor condition was significant ($CFI_{difference} > 0.01$; Cheung & Rensvold, 2002). Furthermore, there was virtually no difference between the fit statistics for free and fixed factor loadings. Moreover, very large standard deviations of NNFI and CFI were observed but not expected for the presence of the difficulty factor condition.

Table 1

Average Fit Indices (SD in italics) Observed in Investigating 500 × 20 Data Matrices Generated With and Without Manipulation According to the Difficulty Factor Condition by the One-factor Confirmatory Factor Analysis Model ($N_{matrices}$ =300)

Difficulty factor condition (DFC)	Type of factor loading	χ^2	RMSEA	SRMR	NNFI	CFI
No DFC	Free	634.6[1]	0.005	0.068	1.001	0.996
		70.6	*0.01*	*0.00*	*0.01*	*0.00*
	Fixed	663.3[1]	0.005	0.074	1.001	0.996
		75.5	*0.01*	*0.01*	*0.01*	*0.01*
DFC	Free	790.9[1]	0.014	0.076	0.935	0.939
		273.9	*0.01*	*0.01*	*0.12*	*0.11*
	Fixed	830.2[1]	0.014	0.084	0.939	0.936
		282.7	*0.01*	*0.02*	*0.11*	*0.11*

[1] χ^2 before robustness correction

Further investigations concentrated on the results obtained by using fixed factor loadings, as in this case variance estimates became available. Figure 2 illustrates the average variance estimates based on all variance estimates. The variance for the absence of the difficulty factor condition is provided on the left and the variance for the presence on the right. As is obvious, the variance drops by 7.6 percent from absence to presence of the difficulty factor condition. In 92.5 percent of the datasets investigated by the tetrachoric correlation-based CFA approach the difficulty factor condition led to a decrease of the variance and in the remaining datasets to an increase. These results were in line with the observation that for 106 out of 300 datasets a substantial CFI difference was observed.

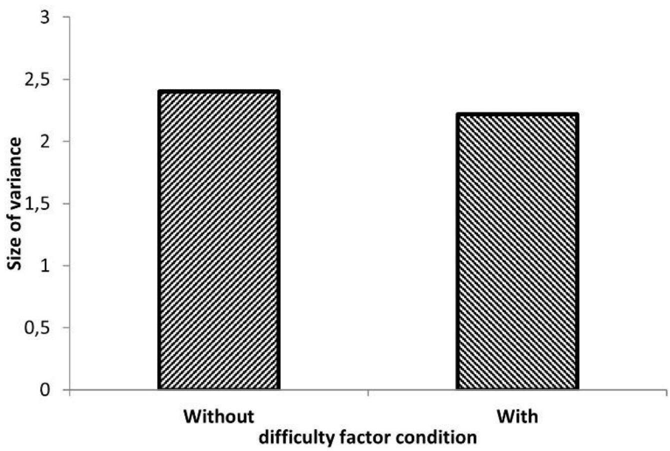

Figure 2. Graphical representation of the average scaled variance estimates of the latent variable in the absence (without) and presence (with) of the difficulty factor condition estimated by a one-factor confirmatory factor model.

Discussion

In the reported factor-analytic investigation the difficulty factor condition is associated with a substantial impairment in model fit when the investigation is conducted by the one-factor CFA model that means without considering a difficulty factor. There is the possibility that this impairment can be avoided by including another factor into the measurement model. There are investigations that demonstrate this possibility (Schweizer & Troche, 2018). Furthermore, there is the decrease of the variance of the latent variable that is predicted based on the deviations observed in investigating the difficulty factor condition. This situation suggests that a difficulty factor is an empirical phenomenon that is closely linked to a specific method effect. It is a method effect that has – unlike other method effects – so far not attracted researchers' attention whereas its consequence for factor analysis, the difficulty factor, stimulated a long-lasting scientific discussion (Hattie, 1985) that started in the 50ties and is still in progress.

General Discussion

We started this essay with highlighting the two major characteristics of a difficulty factor: a relationship between factor loadings and difficulty levels on the one hand and the lack of a specific meaning on the other hand (Guilford, 1941). Another characteristic appears to be the rarity of reports of the observation of such a factor that makes it look more like a mystery or artificial result than a scientifically acknowledged phenomenon (Kubinger, 2003). This lack of scientific credibility may even contribute to the rare reports of such a factor as well as a recently established habit among applied researchers: the use of correlated errors for eliminating disturbing systematic variation. Correlated errors may improve model fit but they do not explain at the expense of which source this is achieved (Schweizer, 2012).

The reported research tries to shed light on the origin of a difficulty factor. For this purpose, we attempted to trace such factors back to method effects. It is well known that method effects can give rise to method factors (Campbell & Fiske, 1959). Method effects due to observational methods or to the context of the application of such methods can influence measurement in a characteristic way, resulting in data that not only demonstrate variation due to the primary source but also additional variation that originates from the observational method or the context of measurement. This additional variation can be considered as modified variation in the sense that it is due to a systematic modification of otherwise created systematic variation.

Furthermore, a method effect can induce a hierarchical structure into data so that subsets of participants differ according to the active influences on performance. For example, speededness induces the creation of two different subsets of participants (Schweizer, Gold, & Krampen, 2020). In participants who are unable to complete an item because of not enough time processing speed determines performance whereas in the other participants it is the source of responding that is stimulated by the item and determines performance (e.g., reasoning).

The results of our investigation show that extreme difficult and equal-sized extreme difficult items which are suggested as sources of a difficulty factor (Bandalos & Gerstner, 2016) lead to deviations of correlations (and covariances) from what is expected. These deviations in turn become apparent in the statistical investigation as additional systematic variation. They amount to the difficulty effect that follows from the difficulty factor condition and characterize the input to factor analysis. This observation suggests that predictability of the difficulty effect is possible. Therefore, given the difficulty effect, a difficulty factor may be predictable. Now researchers conducting factor analysis are better aware of situations demanding the consideration of a difficulty factor than before.

The empirical results also provide insight into a characteristic of the general factor that is expected to account for the systematic variation originating from the stimulated latent source of responding. We learn that the capacity of the general factor to account for the systematic variation shows a restriction. It appears to work well over almost the complete difficulty range of items with the exception of small boundary ranges of extreme difficulty levels. Within these small ranges the general factor (i.e. factor with larger factor loadings of all items) may not account well for systematic variation although there is no other latent source of systematic variation than the difficulty factor condition. This means that the difficulty factor has no substantial source of its own; it is due to a method source only.

The results reported in this study can be useful for future investigations concerning the structural validity of scales. They tell researchers that the presence of items showing extreme difficulty levels increase the probability of a difficulty effect that may necessitate the consideration of a difficulty factor in order to reach good model fit.

References

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). Germany: Hogrefe Publishing.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York: The Guilford Press.

Campbell, D. T., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56(2),* 81–105.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9(2),* 233–255. http://doi.org/10.1207/S15328007SEM0902_5

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(5),* 323–329. http://doi.org/10.1007/BF02288588

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7(3),* 286–299. http://doi.org/10.1037/1040-3590.7.3.286

Gibson, W. A. (1959). Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis. *Psychometrika, 24(3),* 229–252. http://doi.org/10.1007/BF02289845

Gibson, W. A. (1960). Nonlinear factors in two dimensions. *Psychometrika, 25(4),* 381–392. http://doi.org/10.1007/BF02289755

Graham, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement, 66(6),* 930–944. http://doi.org/10.1177/0013164406288165

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60,* 549–576. http://doi.org/10.1146/annurev.psych.58.110405.085530

Guilford, J. P. (1941). The difficulty of a test and its factor composition. *Psychometrika, 6(2),* 67–77. http://doi.org/10.1007/BF02292175

Hattie, J. (1985): Methodological review: assessing unidimensionality of tests and items. Applied *Psychological Measurement, 9(2),* 139–164. http://doi.org/10.1177/014662168500900204

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),* 1–55. http://doi.org/10.1080/10705519909540118

Jöreskog, K. G. (1970). A general method for analysis of covariance structure. *Biometrika, 57,* 239-257. http://doi.org/10.2307/2334833

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*: *User's reference guide*. Lincolnwood, IL: Scientific Software International Inc.

Knowles, E. S. (1988). Item context effects on personality scales: Measuring changes the measure. *Journal of Personality and Social Psychology*, *55*(2), 312–320. http://doi.org/10.1037/0022-3514.55.2.312

Kubinger, K. D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science, 45(1),* 106–110.

Little, T. D., Slegers, D. W., & Card, N. A. (2006). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13(1),* 59–72. http://doi.org/10.1207/s15328007sem1301_3

Lu, Y., & Sireci, S. G. (2007). Validity issues in test speededness. *Educational Measurement, 26*, 29–37. http://doi.org/10.1111/j.1745-3992.2007.00106.x

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology*. http://doi.org/10.3389/fpsyg.2013.00169

McDonald, R. P. (1965). Difficulty factors and nonlinear factor analysis. *British Journal of Mathematical and Statistical Psychology, 18(1),* 11–23. http://doi.org/10.1111/j.2044-8317.1965.tb00690.x

McDonald, R. P. (1967). Non-linear factor analysis. *Psychometric Monograph, 15*. Richmond: Psychometric Cooperation.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. *British Journal of Mathematical and Statistical Psychology, 27(1)*, 82–99. http://doi.org/10.1111/j.2044-8317.1974.tb00530.x

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous variable indicators. *Psychometrika, 49(1),* 115–132. http://doi.org/10.1007/BF02294210

Oshima, T. C. (1994). The effect of speededness on parameter estimation in item response theory. *Journal of Educational Measurement, 31(3),* 200–219. http:// doi.org/10.1111/j.1745-3984.1994.tb00443.x

Satorra, A., & Bentler, P.M. (1994). Corrections to the test statistics and standard errors on covariance structure analysis. In A. von Eye & C. C. Glogg (Eds.), *Latent variable analysis* (pp. 399–419). Sage.

Schweizer, K. (2012). On correlated errors. *European Journal of Psychological Assessment, 28,* 1-2. http://doi.org/10.1027/1015-5759/a000094

Schweizer, K. (2020). Method Effects in Psychological Assessment. *Psychological Test and Assessment Modeling, 64(3),* 337–343.

Schweizer, K., Gold, A., & Krampen, D. (2020). A semi-hierarchical confirmatory factor model for speeded data. *Structural Equation Modeling, 27(5),* 773–780. http://doi.org/10.1080/10705511.2019.1707083

Schweizer, K., & Reiß, S. (2019). On the contextual conditions driving a difficulty factor. *International Journal of Statistics and Probability, 8(5*), 1–12. http://doi.org/ 10.5539/ijsp.v8n5p1

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(1),* 46–69. http://doi.org/10.1177/0013164416670711

Schweizer, K., & Troche, S. (2019). The EV scaling method for variances of latent variables. *Methodology, 15(4),* 175–184. http://doi.org/10.1027/1614-2241/a000179

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. http://doi.org/10.3389/fpsyg.2019.00887

Sechrest, L., Davis, M. F., Stickle, T. R., & McKnight, P. E. (2000). Understanding ‚method‘ varance. In L. Bickman (Ed.), Research design: Donald Campbell’s legacy (pp. 63 – 87). Thousand Oaks, CA: Sage Publications.

Wherry, R. I., & Gaylord, R. H. (1944). Factor pattern of test items and tests as a function of the correlation coefficient: content, difficulty and constant error factors. *Psychometrika, 9(4),* 237–244. http://doi.org/10.1007/BF02288781

Widaman, K. F. (2018). On common factor and principal component representations of data: implications for theory and for confirmatory replications. *Structural Equation Modeling, 25(6),* 829–847. http://doi.org/10.1080/10705511.2018.1478730

Zeller, F., Reiss, S., & Schweizer, K. (2017). Is the item-position effect in achievement measures induced by increasing item difficulty? *Structural Equation Modeling, 24(5),* 745–754. http://doi.org/10.1080/10705511.2017.1306706