VIOLATIONS OF UNIDIMENSIONALITY AND LOCAL INDEPENDENCE IN MEASURES INTENDED AS UNIDIMENSIONAL: ASSESSING LEVELS OF VIOLATIONS AND THE ACCURACY IN UNIDIMENSIONAL IRT MODEL ESTIMATES

by

Gordana Rajlic

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Measurement, Evaluation, and Research Methodology)

THE UNIVERSITY OF BRITISH COLUMBIA
(Vancouver)

July 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Violations of unidimensionality and local independence in measures intended as unidimensional: Assessing levels of violations and the accuracy in unidimensional IRT model estimates

submitted by Gordana Rajlic in partial fulfillment of the requirements for

the degree of Doctor of Philosophy

in Measurement, Evaluation, and Research Methodology

Examining Committee:

Dr. Bruno D. Zumbo, Measurement, Evaluation, and Research Methodology
Supervisor

Dr. Anita M. Hubley, Measurement, Evaluation, and Research Methodology
Supervisory Committee Member

Dr. Amery D. Wu, Measurement, Evaluation, and Research Methodology
Supervisory Committee Member

Dr. Skye Barbic, Rehabilitation Sciences
University Examiner

Dr. Chris G. Richardson, Population and Public Health
University Examiner

Abstract

The current study was motivated by psychological measures intended as unidimensional, in which violations of unidimensionality and local independence (LI) are present. Unidimensionality and LI are the assumptions of unidimensional IRT models, widely used in educational assessment and increasingly used in other fields of psychology and in social and health sciences. Providing more details about the relation between different levels of violations of the two assumptions and the accuracy in unidimensional IRT model estimates was the main goal of the current study. The second goal of the study was to investigate the utility of certain ways to provide recommendations and guidelines about the size of distortions in unidimensional model estimates at different levels of violations. Four indexes based on eigenvalues from exploratory factor analysis were examined for such a purpose. Deemed beneficial for the main purpose of the study, a multidimensional model consistent with a "locally dependent unidimensional model" and a particular research design, based on varying the strength of the relevant latent dimensions, were employed in the study to create conditions of violations of the two assumptions in measures intended as unidimensional. The results of the study demonstrated robustness of the unidimensional IRT model (i.e., 2PL IRT model) under a range of violations and provided more information about the conditions of robustness. A strong relation was demonstrated between the violations, as defined by the strength of local dependence (LD), and the size of the distortions in the item parameters estimation (e.g., correlation of 0.90 with bias in the item discrimination estimates). The item discrimination parameter was systematically overestimated. In relation to the person location parameter, overestimation at the low end of the latent trait and underestimation at the high end of the trait was found – with bias systematically increasing with decrease in the strength of the dominant dimension and with increase in the

strength of LD. Regarding the use of eigenvalue-based indexes in predicting bias, mixed results were obtained (no value was demonstrated in relation to item parameters estimation and some value in relation to person location estimation). The utility of the model/design employed in the study was discussed.

Lay Summary

When individuals respond to items of a measure designed to measure one characteristic, their responses, to a certain degree, reflect other unintended influences in addition to the targeted characteristic. Providing more details about relation between how much items tap into "something else" and the accuracy of the conclusions regarding what we intended to measure was the main purpose of the study. The additional goal of the study was to explore certain ways of forming guidelines for the applied researchers about how great a mistake they might be making in their conclusions by assuming that their measures capture only one characteristic. Simple and convenient indexes, based on the analysis that applied researchers conduct regularly in their work (exploratory factor analysis), were examined. Special consideration was given in the study to the conditions relevant for general psychological measures.

Preface

This dissertation is original, unpublished, independent work by the author, G. Rajlic.

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to thank my research supervisor, Dr. Bruno Zumbo, and the members of my research committee, Dr. Anita Hubley and Dr. Amery Wu, for their thorough and thoughtful review and helpful feedback regarding this dissertation. I gratefully acknowledge the SSHRC Doctoral funding for my research during involvement in the MERM program.

For all the support of my family and their understanding of me, I am forever thankful.

Dedication

To Predrag, Ella, and Thea

1. Introduction

Measures intended to measure a single construct have been used widely in different domains of psychology. Relevant for such measures are unidimensional item response theory (IRT) models (Birnbaum, 1968; Bock, 1972; Lord, 1952; Lord & Novick; 1968; Rasch, 1960; Samejima, 1969). Unidimensional IRT models rest on a set of assumptions, including assumptions of unidimensionality and local independence (LI), which have been characterized as "strong" assumptions. The two assumptions are characterized in such a way as they are difficult to meet in reality; that is, certain violations of unidimensionality and local independence are always present in the real measures (Hambleton & Swaminathan, 1985; McDonald, 2000; Nandakumar, 1991; Stout, 1987; Traub, 1983). Violations of the assumptions may be important or not – this is an empirical question related to the consequences of the violations, such as are, for example, the distortions in parameter estimates (de Ayala, 2009). If at certain levels of the violations the distortions in model estimates are small, unidimensional models still may be appropriate.

The unidimensionality assumption posits that one latent dimension underlies responses on the given measure, whereas LI (or conditional independence) assumes that different items responses are independent conditioning on the underlying trait (Hambleton & Swaminathan, 1985). Local independence actually defines what is meant by unidimensionality (Lord & Novick, 1968; McDonald, 1981); that is, item responses set is unidimensional when item responses are locally independent based on a single latent trait[1]. The two assumptions are distinct concepts;

[1] Local independence can be met in multidimensional models - when multidimensional space is completely mapped and all latent variables accounted for (Lord & Novick, 1968).

however, in the context of the measures intended to measure one construct, violation of one assumption means that the other assumption is, in a certain amount, violated too.

Violations of unidimensionality and different consequences of such violations, in the IRT context, have been addressed in a number of studies (including Ackerman, 1989; Ansley & Forsyth, 1985; Drasgow & Parsons, 1983; Kahraman, 2013; Reckase, 1979; Zhang, 2008). The findings of the majority of the studies pointed to a general conclusion that the closer a set of item responses is to strict unidimensionality, smaller errors in the parameter estimates result from the use of unidimensional IRT models. However, more information about the conditions of robustness of the models is needed. Clear and agreed upon guidelines about how much distortions in the estimates are expected at different levels of violations have not been established (Kahraman, 2013; Ip, 2010). More information about distortions in unidimensional model estimates at different levels of violations would be beneficial in many practical and applied-research contexts. Additionally, with increased interest in the use of IRT models in different fields of psychology and the social sciences (i.e., outside of the educational and achievement assessment/research where they have been traditionally utilized) there is a need for more information about the violations and their consequences in these new contexts (Reise & Rodriguez, 2016).

The general purpose of the current study was to further examine the relation between different levels of violations of the two assumptions and the accuracy of unidimensional IRT models, and to provide more details about the relation. Distortions in estimates of unidimensional model (specifically, binary two-parameter logistic, 2PL, model) were examined at different levels of violations. Of interest were violations of the assumptions more typical for the measures intended as unidimensional (i.e., “weaker” forms of violations) rather than the

violations reflective of a substantive multidimensionality in a measure. In the study, LI violations were modeled in a dimensionality framework (Ip, 2010). A model and a research design, different from the predominantly used ones (as described in the following sections), were employed as potentially beneficial for addressing the main purpose of the study. A measure with a smaller number of items was of interest. The investigation was conducted in the IRT context and the IRT parametrization was used. However, the problems addressed, research methodology, and the results are readily converted into item factor analysis (FA) context and the FA parametrization, based on mathematical equivalence of the two-parameter IRT model for binary responses (used in this study) and FA of item responses (Olsson, 1979; Takane & de Leeuw, 1987; Wirth & Edwards, 2007).

Additionally, as relevant for the problem of violation of the assumptions, certain ways of providing recommendations and forming guidelines about the size of the distortions in unidimensional model estimates at different levels of violations were investigated in the study. In accordance with the employed model/design (described in the following sections), the utility of some of the indicators of the "strength" of latent dimensions (as opposed to the size of correlation between the dimensions, predominantly used in previous research) was investigated in this purpose. Utilizing recommendations provided in terms of "correlation between latent dimensions" is challenging in applied contexts, when the latent dimensions and the correlation between them are not known. The utility of simple and convenient indices of the strength of the latent dimension – eigenvalues from exploratory factor analysis (EFA), was examined in the current study. Previous research in which the indices of the strength of latent dimensions were utilized (in IRT and SEM context) and the research concerned with the assessment/diagnosing of unidimensionality in measures informed the investigation related to the second study purpose

(Bonifay, Reise, Scheines, & Meijer; 2015; Hattie, 1985; Reckase, 1979; Reise, Scheines, Widaman, & Haviland, 2013; Slocum-Gori & Zumbo, 2011; Tate, 2003; Zhang & Stout, 1999). The two interrelated purposes of the study are described in greater detail in the following sections.

1.1 Investigating the Relations between Unidimensionality and Local Independence Violations and the Accuracy of Unidimensional IRT Model Estimates

In the current study, the conditions were generated that depicted different levels of violations of unidimensionality and LI, in measures intended as unidimensional, and the accuracy in the unidimensional IRT model (2PL model) estimates were then assessed in the created conditions. To depict the violations characteristic of measures intended to measure one construct, a model consistent with a "locally dependent unidimensional model" (Ip, 2010) was used. In this model, instead of single dimension underlying item responses, it is posited that a major, dominant dimension underlies the responses, with some dependency among item responses remaining after controlling for the target trait (consistent with general concept of essential unidimensionality; Humphreys, 1964; Nandakumar, 1991; Stout, 1987). Ip (2010) demonstrated that locally dependent unidimensional models and multidimensional models are empirically indistinguishable. Therefore, in the current study, by utilizing multidimensional IRT model (MIRT; Reckase, 2009), data structures were simulated with two underlying latent dimensions: a) a dominant dimension, representing the target construct that the measure was intended to measure – with all items primarily tapping into this dimension, and b) a secondary, nuisance dimension representing source(s) of item dependencies remaining after the construct of interest was accounted for (i.e., local dependence in a measure, LD).

In accordance with the study design, the strength of the two simulated dimensions was systematically varied. Following the FA tradition, the strength of the latent dimensions was defined in the study by the strength of the relations between the items and the dimensions (i.e., by magnitude of item discriminations in IRT parametrization or factor loadings in FA parametrization). The magnitude of item discrimination on the two dimensions was varied and several combinations of the strength of the dimensions were created, with the secondary dimension of less strength in all of the conditions compared to the dominant one. For example, condition 1 (strong – negligible) was a combination of a very strong dominant dimension ('very high' item discrimination on that dimension) and negligible secondary dimension ('very low' item discrimination on that dimension), as described in Simulated conditions section.

Based on varying the strength of the two dimensions, ten conditions were created that depicted different conditions of violations of strict unidimensionality and LI in measures intended as unidimensional. After applying unidimensional IRT model in the created conditions, the accuracy of the model estimates regarding the target construct (i.e., the dominant dimension) was then assessed. Specifically, bias and root mean square error (RMSE) were examined in the estimates of item parameters (item location and item discrimination) and of person location parameter. The contribution of the two factors – the strength of the LD and the strength of the dominant dimension, to the distortions in the unidimensional model estimates was explored.

The model used for creating the desired conditions of violations of the assumptions, was an exploratory MIRT model, in which each item loads on more than one simulated dimension. That is, following a newer line of unidimensional models robustness research (Kahraman, 2013; Sahin, Walker, & Gelbal, 2015; Zhang, 2008), complex item responses were utilized. Complex items have been described as items that do not share one common attribute only (McDonald,

1999); in addition to reflecting the target construct, the item responses reflect something else as well. When items are restricted in a measurement model to discriminate on only one of the dimensions, such items are characterized as "simple-structure items". Simple-structure items have been utilized more often in the IRT robustness research compared to complex items and the related "within-item" multidimensionality[2]. The concept of complex items responses, at the same time, is more realistic and relevant for measures in various scientific domains. In psychology, for example, due to the factors related to the nature of phenomena of interest (i.e., their complexity and interrelatedness), heterogeneity in the respondents' populations, and complexities in the testing contexts, items that reflect one construct only are difficult to find.

What, exactly, the items tap into, in addition to the construct they were designed to measure, can be difficult to pinpoint. In relation to this, the nature of violations of unidimensionality in real–life measures, intended to measure a single construct, has been debated often (Alessandri, Vecchione, Eisenberg, & Łaguna, 2015; Marsh, Scalas, & Nagengast, 2010, Slocum-Gori, Zumbo, Michalos & Diener, 2009). Therefore, in the current study, a secondary dimension that the items tap into was conceptualized in a very general way - as residuals, "the item dependencies not accounted for by the dominant dimension", without focusing on a specific source(s) and reason(s) for such item dependencies. In this way, a more general context for studying violations of the assumptions and their consequences was provided, compared to the

[2] Within-item multidimensionality refers to multidimensional model in which individual items load on more than one dimension. The designs for studying violation of unidimensionality based on use of simple-structure items are characterized as "between-item" multidimensionality – each item loads on only one dimension in such models.

studies concerned with specific types of LD[3]. Such conceptualization and the model utilized in the study may be beneficial when the focus is on studying general relations and establishing general rules that could be relevant for a wider range of measures/measurement situations.

1.2 Investigating Utility of Eigenvalues-Based Indexes in Predicting Distortions in IRT Models Estimates

An additional, interrelated purpose of the current study was to further explore the ways to quantify the relation between the amounts of unidimensionality/LI violations and the distortions in unidimensional model estimates. Specifically, the eigenvalues from principal-axes factoring in EFA (Gorsuch, 1983; Mulaik, 2009) were investigated, with the purpose of assessing if they could be used in forming guidelines about the amounts of distortions at different levels of the violations. Eigenvalues provide information about the variance accounted for by the latent dimensions; therefore, they are a convenient choice for expressing the strength of the latent dimensions. In the current study, EFA was conducted with the data based on the different conditions of violations of unidimensionality/LD (created by manipulating the strength of the simulated latent dimensions, as described in the previous section), and eigenvalues were computed. Based on eigenvalues, several indexes were formed and the utility of the eigenvalues-based indexes was then investigated in predicting bias, and in forming the guidelines about the distortions in unidimensional model estimates, at different levels of violation of

[3] For example, LD due to items taping in other related construct(s) was often studied. In such situations, a model with correlated dimensions (correlated to a certain degree, depending on what specific constructs are in question) would be appropriate ('correlated-traits model'). For some of the sources of LD in measures see Steinberg and Thiessen (1996) and Yen (1993), and for different ways of modeling LD see Chen and Thissen (1997) and Ip (2010). For the conceptualization of the LD in this study, a model that corresponds to "unrotated" EFA/MIRT model, with uncorrelated dimensions, is appropriate.

unidimensionality and LI. In other words, the utility of eigenvalues/eigenvalue-based indexes as possible 'diagnostic indices' in relation to the consequences of the violations was investigated.

The investigation of utility of eigenvalues in predicting bias was motivated by previous research about different ways of indexing unidimensionality in measures (Reise et al., 2013a; Rodriguez et al., 2016; Stout, 1987; Zhang & Stout, 1999), and previous research investigating the use of eigenvalues in assessing dimensionality/deciding how many factors to retain in FA solutions. Various methods for deciding about the number of factors to retain in EFA were based on the use of eigenvalues, such as 'eigenvalue greater than 1' rule, scree plot, parallel analysis (Cattell, 1966; Horn, 1965; Kaiser, 1960). The utility of eigenvalues was most often investigated in such a context (Ruscio & Roche, 2012; Slocum-Gori & Zumbo, 2011; Velicer, Eaton, & Fava, 2000; Yang & Xia, 2015; Zopluoglu & Davenport, 2017).

One of the possible benefits of use of eigenvalues/eigenvalue-based indexes as diagnostic in deciding beforehand whether the unidimensional IRT modeling is appropriate or not, is their convenience. Due to their computational simplicity and intuitive interpretation, eigenvalues could be more convenient for applied researchers to obtain and use, compared to other measures recommended in IRT context, such as DETECT (Zhang & Stout, 1999) and DIMTEST (Nandakumar & Stout, 1993; Stout, 1987, 1990), or the measures proposed in SEM context, based on bifactor models (Rodriguez et al., 2016; Reise et al., 2013a). The use of eigenvalues has been proposed as possibly beneficial in assessing the assumptions of IRT models (Ackerman, 1989; Embretson & Reise, 2000; Hambleton & Murray, 1983; Lord, 1980); however, empirical evidence about the relations between eigenvalues and the consequences of the assumptions violations is limited.

It should be highlighted that the current study was conceptualized from a latent variable framework (i.e., common-factors framework), therefore the eigenvalues were computed from the correlation matrix with the communalities on the diagonal (as in principle-axes factoring), not from the correlation matrix with '1's on the diagonal (used in principal component analysis), as further elaborated in the Method section. To index different levels of violations of undimensionality/LI in the current study, four eigenvalues-based indexes were formed and the relations between the indexes and bias in the unidimensional IRT model estimates were then investigated. The four selected indexes were: Index 1 – the ratio of first to second eigenvalue; Index 2 – the difference between first and second eigenvalue; Index 3 – the proportion of overall standardized variance in a measure accounted for by the first common factor; and Index 4 – the proportion of overall common variance in a measure accounted for by the first common factor.

Index 1 and index 2 reflect the strength of the first two common factors in FA solution, relative to each other. They are based on the first and second eigenvalues only[4]; specifically, Index 1 is a ratio of, whereas Index 2 is a difference in the first two eigenvalues. Index 3 reflects the strength of the first common factor relative to the overall standardized variance in a measure[5]. It is ratio of the first eigenvalue and the number of items in a measure (which is 15 in this study). Index 4 reflects the strength of the first common factor relative to the overall common variance in a measure. It is ratio of the first eigenvalue and the sum

[4] The second eigenvalue is the largest of all other eigenvalues (or possibly equal but not smaller than the other eigenvalues).

[5] Overall standardized variance in a measure is a sum of the elements on the diagonal of the inter-item correlation matrix (original correlation matrix with 1s are on the diagonal). It equals the number of items in a measure (i.e., 15 in this study).

of all eigenvalues[6]. These four indexes were selected as they, or similar indices, were either proposed as useful, or were utilized, in the previous research concerned with assessing dimensionality of measures or with assessing degree of unidimensionality in measures (Reise et al., 2013a; Rodriguez et al., 2016; Sijtsma, 2009; Slocum-Gori & Zumbo, 2011).

[6] Overall common variance in a measure is a sum of the elements on the diagonal of the correlation matrix with communalities on the diagonal. It was less than 15 in the created conditions, as the values less than 1 were on the diagonal of the correlation matrix. Whereas the overall standardized variance was same in all created conditions, the overall common variance varied in the conditions.

2. Literature Review and Background

Violations of unidimensionality have been studied in several studies, mostly conducted with simulated data (Ackerman, 1989; Ansley, & Forsyth, 1985; Crişan, et al. 2017; Drasgow & Parsons, 1983; Folk & Green, 1989; Harrison, 1986; Kahraman, 2013; Kirisci, et al., 2001; Reckase, 1979; Reise et al., 2013b; Sahin, et al, 2015; Zhang, 2008). There was a substantive heterogeneity in the conducted studies in several aspects, including what specific research questions were asked and what methodology was used in addressing them. For example, in relation to the research question, in some studies a question of primary interest was what it is that the estimates from unidimensional models actually estimate when the model is multidimensional (Ansley & Forsyth, 1985; Way, Ansley, & Forsyth, 1988). In the other studies, the focus was placed on specific consequences of violations of unidimensionality. When consequences of violations of unidimensionality were of interest, the most commonly studied consequences were distortions in model estimates (i.e., distortions in item parameters and/or person parameter estimates), whereas some other consequences studied were those related to specific practical applications, such as consequences for adaptive testing item selection (Folk & Green 1989), or for ranking of individuals and selection decisions (Crişan et al., 2017). In terms of data generation procedure – in some studies the data were generated utilizing FA framework (Drasgow & Parsons, 1983; Harrison, 1986; Reckase, 1979), whereas in the others, IRT techniques were used (Kahraman, 2013; Zhang, 2008).

Important differences in the studies include how unidimensionality violations were conceptualized, what model was used to represent violations, and how undimensionality violations were manipulated. Multidimensionality models such as second-order models (Drasgow & Parsons; 1983; Harrison, 1986), bifactor models (Bonifay et al., 2015, Reise et al.,

2013b), as well as multidimensional models without hierarchical structure (several correlated factors in Ansley & Forsyth, 1985; Kirisci, et al., 2001; and a dominant and some minor dimensions in Kahraman, 2013 and Zhang, 2008) were used. In terms of how dimensionality (i.e., violation of unidimensionality) was manipulated, it was predominantly done by manipulating the correlation between the latent dimensions, generated by a specific multidimensional model. In some of the studies, the violations were manipulated by varying proportion of items loading on more than one dimension, in addition to manipulating the correlations, or by varying the amount of 'misfitting' items.

In relation to the estimation procedure, different estimation methods and estimation software were utilized in the studies, with LOGIST (based on joint maximum likelihood estimation; Wingersky, Barton, & Lord, 1982), and BILOG (based on marginal maximum likelihood estimation, MMLE; Mislevy & Bock, 1983) used the most frequently. In terms of the type of unidimensional IRT model that was applied to the generated data 2PL model was utilized the most often. About the type of parameter estimates of interest, some of the studies focused on the accuracy of item parameter estimates, others on theta estimates, and some investigated both.

Taking into account the heterogeneity in conceptualizations of violations of the assumptions, it is difficult to conveniently summarize the results and the conclusions of the studies (Kirisci, et al., 2001). In the most general terms, the conclusion that could be formulated concerning robustness of unidimensional IRT models, based on the majority of the results of the conducted studies, is that as closer the data structure was to strict unidimensionality, smaller error in the parameter estimates resulted from the use of unidimensional IRT models. "Closeness" to strict unidimensionality, however, was defined and assessed in different ways in different studies; hence, the results and recommendations were formulated in different ways.

When the focus of research was on violations of unidimensionality consistent with multiple related dimensions/factors, magnitude of the correlation between the dimensions was used in providing recommendations in relation to robustness of the models. In Kirsci et al. (2001), based on three-dimensional data structures, the authors recommended that if the latent dimensions were highly correlated (r >.40), and the correlations were approximately the same across dimension pairs, the application of unidimensional IRT models was feasible. Similar recommendation resulted from the studies that utilized two-dimensional data (Ackerman, 1989; Folk & Green, 1989). However, in two similar studies with two latent dimensions generated (Ansley & Forsyth, 1985; Way et al. 1988), somewhat different results were obtained; that is, even at the high level of correlation between the latent dimensions, the obtained estimates were substantially different from the corresponding true parameter values.

The results of the studies that focused on violations consistent with the multiple correlated dimensions but employed a hierarchical factor model (Drasgow & Parsons, 1983; Harrison, 1986) also provided recommendations in terms of the correlations among the first order factors. The item parameters estimates and latent trait estimates were highly correlated with factor loadings and factor scores for general factor when the first order factors were highly correlated (r = .46 to .90). In the studies in which closeness to strict unidimensionality was measured by existence of a strong dominant factor, the percentage of variance accounted for by the dominant dimension was used in summarizing the results (Reckase, 1979). The findings in Reckase (1979), based on real and simulated data, suggested that the size of the first factor controlled the estimation of the IRT parameters (such as the size of average 3PL model discrimination parameter estimates, stability of the 3PL model difficulty estimates, 1PL model probability of fit, and the mean squared deviations for both 1PL and 3PL models). The author

concluded that good latent trait estimates and acceptable item calibration were obtained when the first factor accounted for at least 20 percent of the test variance.

In other studies in which closeness to strict unidimensionality was represented by existence of a strong dominant dimension, proportion of the items that load on the dominant dimension, and correlations between the dominant and minor dimensions, were used in summarizing results and providing recommendations. According to Zhang (2008), when simulated multiple secondary dimensions were highly correlated with the dominant dimension, and small/medium number of items measured the secondary dimensions, unidimensional models generally fit, and the accuracy of latent trait estimation was comparable to that of strictly unidimensional tests. In Kahraman (2013) and Sahin et al. (2015), the discrimination parameter estimates of the complex-structure items were subject to various degrees of estimation bias, increasing in severity as the correlations between the primary and secondary traits *increased*, and the number of complex-structure items increased. Item difficulty parameter estimates were reasonably free from bias (Kahraman, 2013); whereas in relation to the latent trait estimate, errors decreased as the correlation between dimensions increased (Sahin et al., 2015). Crişan et al. (2017) varied the proportion of misfitting items and correlation between two latent dimensions - the authors concluded that as the proportion of misfitting items increased and correlation between dimensions decreased, precision of the latent trait estimate decreased.

In relation to the studies that focused on manipulating correlations between the factors/dimensions, and provided guidelines about the consequences of the unidimensionality violations based on the magnitude of correlations between the dimensions, it is problematic that in practical/applied contexts, latent dimensions as well as correlations between them are unknown. Making decisions about dimensions underlying the data and estimating correlations

between them is a process that can be highly subjective (e.g., dependent on the choice of method of factor rotation). Consequently, using correlations between latent dimensions as a criterion, very different conclusions could be made about possible distortions in unidimensional model estimates, based on the same data. Other means that could be used for providing recommendations regarding violations of unidimensionality and their consequences, which may be less prone to individual interpretation, are the indices recommended in the context of IRT model-fit checking, and those proposed in the context of bifactor modeling.

In the IRT context, DETECT index has been recommended for assessing essential unidimensionality. Empirical evidence about the utility of DETECTS in predicting bias and in providing recommendations about consequences of unidimensionality/LI violations is lacking. A limited utility of DETECT index in predicting bias in the unidimensional model estimates, in SEM context, was suggested in Bonifay et al. (2015). The authors recommended using DETECT in combination with other measures of general factor strength, specifically, the percentage of uncontaminated correlations (PUC) and the explained common variance (EVC). The indices based on bifactor models, such as PUC, omega reliability coefficients (omega, omegaS, omegaH, omegaHs), and factor determinacy, have been proposed as useful in indexing unidimensionality/essential unidimensionality (Bonifay et al., 2015; Reise et al., 2013a; Reise et al., 2013b; Rodriguez, et al., 2016). More empirical evidence about value of such indices, for the purpose of providing recommendations regarding violations of unidimensionality/LI and their consequences, in the IRT context, is needed.

As possibly beneficial in assessing unidimensionality violations in IRT context, eigenvalues from PCA and EFA have been proposed (Ackerman, 1989; Embretson & Reise, 2000; Hambleton & Murray, 1983; Lord, 1980). Eigenvalues are simpler compared to the

methods proposed in the IRT model-fit and bifactor contexts; that is, they are easy to obtain and interpret by practical researchers and, as such, they could be handy tools in providing guidelines about the amounts of violations of the two assumptions and their consequences. Eigenvalues and eigenvalues-based indexes, however, have been rarely utilized in the IRT robustness research. Ackerman (1989) provided information about the ratio of first to second eigenvalue in his study - the author reported that as the correlation between latent traits increased, the ratio of first to second eigenvalue increased, indicating more dominant first principal component. At correlation of 0.90 data were almost unidimensional, with the ratio of first to second eigenvalue approaching 7. Reckase (1979) used the size of the first eigenvalue (i.e., the proportion of variance accounted by the first factor) in formulating the conclusions about the robustness of the model used in this study. In general, indices based on eigenvalues have been more often used in the context of diagnosing unidimensionality of measures, and their utility has been primarily studied in that context (Hattie, 1985). Such use of eigenvalues has not been reviewed here, as the purpose of the use of eigenvalues was different in this study, that is, the goal was studying the utility of eigenvalues and eigenvalues-based indexes in providing guidelines about the unidimensionality/LI violations and their consequences.

Finally, in a different line of research concerned with robustness of unidimensional IRT models to the assumptions violation, the relevant issues were addressed from the perspective of LI (Chen & Thissen, 1997; DeMars, 2012; Ip, 2010; Sireci et al., 1991; Wainer et al., 2007; Wainer & Thissen, 1996; Wang et al., 2005; Yen, 1984, 1993). This research was conducted mostly in in the context of educational/achievement assessment and measurement and with longer measures. Similarly as in the research concerned with violations of unidimensionality, there was a substantive heterogeneity in the research about violations of LI - in respect to the

way that violations of LI were conceptualized (i.e., what specific type of LD was of interest) and in methodology used (what measures of LI were used and what consequences were assessed). LD was commonly measured at item-pairs level, using measures such as Q_3, G^2, X^2 (Chen & Thissen, 1997; Yen, 1984, 1993). The results of the studies, overall, indicated negative effects of LD in measures and consequences such as an overestimation of test information and model fit, and in bias in item parameters.

As presented in the current chapter, there has been a significant variety in the ways violations of unidimensionality and LI were conceptualized and studied. The conducted research yielded important information about violations of the two assumptions and their consequences in many specific contexts. However, despite a large body of research, further information about the relevant relations and conditions of robustness are needed. That is, greater details about "general" relation between different levels of the violation and the size of possible error in unidimensional models estimates (estimates of item discrimination, item location, and person location) are still needed. Additionally, considering the limitations of using the magnitude of correlations between latent traits in forming such rules, investigation of different ways is needed. The current study has been designed to provide more information in these directions; that is, a model/research design has been employed in the study that may be beneficial for such a purpose.

3. Method

The two interrelated purposes set in the current study were addressed in a Monte Carlo simulation study, following the guidelines provided in Harwell, Stone, Hsu, and Kirisci (1996) and Feinberg and Rubright (2016). A number of data sets (binary item responses) were simulated corresponding to the conditions designed in the study. One of the advantages of a simulation study is that true parameters are known, and that model-estimated parameters can be compared to the true parameters. In the current study, the estimated parameters were estimated item location, item discrimination, and person location obtained by applying the unidimensional 2PL IRT model to the simulated item responses. The true parameters, with which the estimated parameters were compared, were those used in the item responses simulation (specifically, the values corresponding to the *dominant* dimension in the simulated conditions, representing the construct that the measure is intended to measure). The discrepancy between the true and estimated parameters was then recorded, in accordance with the study purposes. Replication design was used in the study, with 100 replication in each of the study conditions. Estimated bias in item discrimination, for example, in each of the conditions, was mean bias across 100 replications.

3.1 Model

The model utilized in the study was a multidimensional model (specifically, two-dimensional); however, it was not intended as a meaningful multidimensional model. Rather it was conceptualized as a unidimensional model with a certain amount of LD, overall consistent with the concept of "locally dependent unidimensional model" (Ip 2010). LD was defined broadly in the current study as "item dependencies not accounted by the dominant dimension", that is, without positing a specific source/reason behind such dependencies. Hence, LD was

represented by a secondary, nuisance dimension (i.e., residuals), uncorrelated to the dominant one. This was done in order to provide a more general framework for investigating the issues of interest. It could represent various real-life measures and measurement contexts, when different sources of LD are relevant. Furthermore, only one secondary dimension was postulated in the model, representing all item dependencies not accounted for by the dominant dimension. One secondary dimension was used as a proxy for possible multiple secondary dimensions, that is, for possibly multiple sources of LD. Additionally, the model could be characterized as "within-item multidimensional" as opposed to "between-item multidimensional" model.

The sources of LD in a measure, modeled in dimensionality framework, could be represented by various multidimensional models, such as multidimensional models with several dimensions/several secondary dimensions[7], and hierarchical/bifactor models. The model with one secondary, nuisance, dimension is simpler than the other models and, due to its simplicity, it could be beneficial in the research contexts, such as the current study, in which further distinguishing among possible different sources of LD was not of primary interest.

3.2 Simulated Conditions

In relation to the first purpose of the current study, the conditions of different amounts of violations of the unidimensionality/LI were created by manipulating two elements in the model – the strength of the dominant dimension and the strength of the secondary dimension, so that the secondary dimension was always of less strength compared to the dominant dimension. Four levels of strength of the dominant dimension ('very strong', 'strong', 'moderate', and 'weak')

[7] Difficulties in estimating MIRT models with higher dimensions were pointed to in Ip & Chen, 2012; Reckase, 2009.

were crossed with four levels of strength of the secondary dimension ('strong', 'moderate', 'weak' and 'negligible'). These levels of strength of the dimensions were characterized by different magnitude of item discriminations, following the guideline about the magnitude of item discrimination in IRT provided by Baker (2001): very high item discrimination ($a > 1.70$), high item discrimination (a range 1.35 - 1.69), moderate item discrimination (a range 0.65 - 1.34), low item discrimination (a range 0.35 - 0.64), and very low item discrimination (a range 0.01 - 0.34)[8]. Given that in the study design the secondary dimension was conceptualized as always of less strength compared to the dominant one, not all pairings of the strength of the dominant and secondary dimensions were plausible. Hence, a "partially-crossed" research design was used, and ten conditions of interest were created (Figure 1). The ten conditions are characterized by the following combinations of the item discriminations magnitude:

1. Very strong – negligible: a very strong dominant dimension with a negligible secondary dimension, characterized by very high item discriminations on the first dimension ($a > 1.70$) and very low item discrimination on the second dimension (a range 0.01 - 0.34);
2. Very strong – weak: a very strong dominant dimension with a weak secondary dimension, characterized by very high item discriminations on the first dimension ($a > 1.70$) and low item discrimination on the second dimension (a range 0.35 - 0.64);

[8] Logistic model item discrimination parameter values are provided in Baker's (2001) classification. In FA parametrization (following Wirth & Edwards, 2007), very high item discrimination in Baker's classification corresponds to factor loadings (FL) from categorical confirmatory FA of .71 and greater; high item discrimination corresponds to FL range .62 - .70; moderate item discrimination corresponds to FL range .36 - .61; low item discrimination corresponds to FL range .20 - .35; and very low item discrimination corresponds to FL range of .01 - .19.

3. Very strong – moderate: a very strong dominant dimension with a moderately strong secondary dimension, characterized by very high item discriminations on the first dimension (a > 1.70) and moderate item discrimination on second dimension (a range 0.65 - 1.34);
4. Very strong – strong: a very strong dominant dimension with a strong secondary dimension - characterized by very high item discriminations on the first dimension (a > 1.70) and with high item discrimination on second dimension (a range 1.35 - 1.69);
5. Strong – negligible: a strong dominant dimension with a negligible secondary dimension, characterized by high item discriminations on the first dimension (a range 1.35 - 1.69) and very low item discrimination on the second dimension (a range 0.01 - 0.34);
6. Strong – weak: a strong dominant dimension with a weak secondary dimension, characterized by high item discriminations on the first dimension (a range 1.35 - 1.69) and low item discrimination on the second dimension (a range 0.35 - 0.64);
7. Strong – moderate: a strong dominant dimension with a moderately strong secondary dimension, characterized by high item discriminations on the first dimension (a range 1.35 - 1.69) and moderate item discrimination on the second dimension (a range 0.65 - 1.34);
8. Moderate – negligible: a moderate dominant dimension with a negligible secondary dimension, characterized by moderate item discriminations on the first dimension (a range 0.65 - 1.34) and very low item discrimination on the second dimension (a range 0.35 - 0.64);
9. Moderate – weak: a moderate dominant dimension with a weak secondary dimension, characterized by moderate item discriminations on the first dimension (a range 0.65 - 1.34) and low item discrimination on the second dimension (a range 0.35 - 0.64);

10. Weak – negligible: a weak dominant dimension with a negligible secondary dimension, characterized by low item discriminations on the first dimension (*a* range 0.35 - 0.64) and very low item discrimination on the second dimension (*a* range 0.01 - 0.34).

Partially-crossed research design was utilized in the study as deemed more relevant for studying violations in measures intended as unidimensional, in which the most likely are weaker forms of violations. Confounding between the effects of the weaker and stronger forms of multidimensionality is avoided by using such a design.

The ten designed conditions, listed above, were preliminarily ordered by the strength of the dominant dimension in the first place and the strength of the secondary dimension in the second place (and numbered accordingly), with the contribution of the strength of the two dimensions to the distortions in model estimates to be investigated in the study. A set of item responses was simulated for each condition (for 5000 respondents), using a MIRT model and parameters values described in the sections about data generation. As multiple replications are recommended and are a standard in IRT simulation studies (Harwell et al., 1996; Feinberg & Rubright, 2016), one hundred replications of each cell in the research design were conducted (i.e. 100 item responses sets were generated in each condition). In this way, a sampling distribution of the model estimated parameters could be studied.

DOMINANT DIMENSIONS STRENGTH

SECONDARY DIMENSION STRENGTH	VERY STRONG	STRONG	MODERATE	WEAK
NEGLIGIBLE	Very Strong/ Negligible	Strong/ Negligible	Moderate/ Negligible	Weak/ Negligible
WEAK	Very Strong/ Weak	Strong / Weak	Moderate/ Weak	
MODERATE	Very strong/ Moderate	Strong/ Moderate		
STRONG	Very strong/ Strong			

Figure 1. Ten simulated conditions of violations of unidimensionality and local independence. The conditions were created by combining four levels of strength of the dominant dimension and four levels of strength of secondary dimension, with the dominant dimension always stronger than the secondary dimension. 100 sets of item responses were simulated for each condition.

It should be highlighted that the ten conditions are primarily "experimental" conditions (i.e., result of systematic varying of the research variables' levels), designed to investigate the relations between the relevant variables and the consequences of interest (e.g., bias in unidimensional model estimates). Systematic varying of the variables of interest could not be achieved with the real measures. The ten simulated conditions were not intended to represent the individual real-life measures, and they are different from the individual real-life measures used in

practice in some important aspects. Some of the conditions depict violations relevant for a range of well-designed measures (i.e., the conditions simulated with a strong dominant dimension and a negligible/weak secondary dimension are similar to the latent structure of the measures, intended to capture a single construct, with one strong dominant dimension and some weak local item dependencies). Other conditions are more extreme, unlikely to be found in well-designed measures, such as the condition with a weak dominant dimension and a negligible secondary dimension. All of the simulated conditions, however, are in certain ways different from the real-life measures, such as in variability of the simulated item parameters and number of the secondary dimensions, which will be discussed further in the section about item parameters simulation, and in relation to the generalizability of the results of the study.

3.3 Item Parameters Generation

For the ten conditions representing violations of unidimensionality and LI, item location parameters ('*b*' parameters) for the two dimensions were generated randomly from the standard normal distribution, as commonly done in the IRT simulation studies (Feinberg & Rubright, 2016). The same '*b*' parameters were used across the ten conditions, in order to focus on the effects resulting from the experimental manipulation (manipulation of '*a*' parameters on the two dimensions), and minimize any other sources of variability across the conditions. In terms of item discrimination parameters ('*a*' parameters), they were prescribed by the study design, based on Baker's (2001) classification of item discrimination strength, as outlined in the description of the ten created conditions. Specifically, '*a*' parameters were simulated from a normal distribution in such a way that '*a*' values for the majority of the items in each designed condition belonged to range of '*a*' values provided by Baker's classification. The mean and standard deviation of the '*a*' parameter distribution were chosen accordingly. For example, by Baker's classification 'low'

item discrimination covers the '*a*' range from 0.35 to 0.64; hence, for a dimension with low item discrimination (i.e., 'weak' dimension), '*a*' parameters were generated from the normal distribution with a targeted mean of 0.50 and standard deviation of 0.15. In this way, the majority of the items[9] had '*a*' in the range 0.35 to 0.64 while allowing some items to have '*a*' values outside of the prescribed range. A guide for simulating '*a*' parameters for dominant and secondary dimensions in the ten created conditions is provided in Table 1.

The same general principle was followed in '*a*' parameters simulation across the ten conditions, with slight adjustments needed in few instances, in order to either stay within the mathematical bounds of FA[10], or to follow the specifics of the research design[11]. As a result, small deviations occurred of the mean and SD for the '*a*' parameters from the intended mean and SD. The item parameters simulated for one of the conditions (specifically, condition 1) are presented in the Table 2, whereas the item parameters for each of the other conditions are enclosed in the Appendix A.

The central aspect of this study was to vary the strength of the dimensions across the designed conditions in a systematic way. The values of '*a*' parameters could have been chosen in

---

[9] Approximately 70% of items, as approximately 70% of cases fall within $M \pm 1SD$

[10] The magnitude of total variance (and communality) in items could not exceed 1, which was particularly of relevance for condition 4.

[11] For example, item discriminations were limited to positive values, so the negative values obtained in random parameter generation for 'negligible' dimension were changed into positive values. In one case, when a cross-loading not allowed by the research design was obtained, the randomly generated parameter was slightly changed according to the predetermined rule. The open ended interval $a > 1.7$ for 'very high' item discrimination was changed into $1.7 < a < 2.1$ in the parameters generation. Additionally, for estimation reasons, 'a' parameter for one of the items on the secondary dimension was set to 0 in all conditions (de Ayala, 2009), resulting in slightly different *M* and *SD* for the 'a' parameters on the secondary dimension compared to *M* and *SD* when the same strength was simulated for the dominant dimension.

different ways as long as their systematic manipulation was possible. For example, all items in one condition could have been assigned the same '*a*' value, which then would be varied across the conditions. The way chosen for generating '*a*' parameters in the current study represents an attempt to make the conditions somewhat more realistic compared to the conditions in which all items would have the same item discrimination value, or compared to generating from the uniform distribution with minimum and maximum that correspond to the values prescribed by Baker's classification. By allowing certain amount of items to have '*a*' values outside of the prescribed ranges, more variability in the generated parameters was provided. However, variability in '*a*' in the designed conditions was still more restricted in comparison to the values found in the real-life measures. Examples of the IRT discrimination in the real-life measures can be found in Embretson and Reise (2000) and de Ayala (2009). Restricting variability in '*a*' was needed in order to systematically vary strength of the two clearly defined dimensions while not allowing for certain types of cross-loadings, i.e., the loadings larger on the secondary dimension than on the dominant one, more characteristic of true multidimensional measures. Such type of cross-loadings were avoided in the current study as the focus was on the conditions consistent with measures intended/designed as unidimensional. Limiting '*a*' to positive values was done for the same reason.

In terms of the type of design for item parameter generation, the same sets of '*a*' parameters were used across the designed conditions, for example, the same set of '*a*' parameters for the 'very strong' dimension was used for the dominant dimension in the conditions 1, 2, 3 and 4. This was done in order to keep the focus on variability resulting from experimental manipulation across the conditions (manipulation of the '*a*' parameters on the two dimension) while minimizing variability resulting from random selection of the parameters within designed

conditions. Similarly, the same two sets of '*b*' parameters were used in all conditions, in order to focus on the variability due to experimental manipulation and minimize variability from any other sources. The two sets of '*b*' parameters were generated from the two successive independent draws from the normal distribution ('*b*' parameters and the histograms of their distribution are included in the Appendix A). While item parameters ('*a*' and '*b*') were treated as fixed across the replications (i.e., the same item parameters are used across replications for each of the conditions), the person location parameter was treated as a "random effect" (i.e., new parameters were generated in each replication from a standard normal distribution). Person location parameter as a random effect is characteristic for the IRT marginal maximum likelihood estimation of item parameters (de Ayala, 2009).

About item format, dichotomous items are utilized in the current study. Dichotomous items have been predominantly utilized in the previous IRT robustness research. Even though more basic, such items are suitable for studying many psychological phenomena and they have been used in psychological measures in various domains[12]. Whereas the focus in the current study is on dichotomous items, the research questions addressed in the study should be, in future, investigated with the polytomous items. In terms of number of items, 15 items were simulated in the current study. In the previous studies about robustness of unidimensional IRT models, primarily concerned with measures in education, the length of measures typically ranged from 15 to 60 items (Crişan et al., 2017). The shorter length was chosen in the current study consistent

[12] Some of the measures that utilize dichotomous item scoring format are the Eysenck's Impulsivity Inventory, the Beck Hopelessness Scale, the Geriatric Depression Scale, the Scales of Psychological Well-Being, the Self-Report Delinquency scale, the Mood Disorder Questionnaire, the Schizotypal Personality Questionnaire, the scales of the Minnesota Multiphasic Personality Inventory, etc.

with the general trend towards shorter measures in psychology, when measures are intended as unidimensional[13].

Table 1

A Guide for Simulating Item Discrimination Parameters for the Ten Designed Conditions

Designed Conditions	a_1 *Mean (SD)*	a_2 *Mean (SD)*
1. Very strong - Negligible	1.90 (0.20)	0.18 (0.17)
2. Very strong - Weak	1.90 (0.20)	0.50 (0.15)
3. Very strong - Moderate	1.90 (0.20)	1.00 (0.35)
4. Very strong - Strong	1.90 (0.20)	1.52 (0.17)
5. Strong - Negligible	1.52 (0.17)	0.18 (0.17)
6. Strong - Weak	1.52 (0.17)	0.50 (0.15)
7. Strong - Moderate	1.52 (0.17)	1.00 (0.35)
8. Moderate - Negligible	1.00 (0.35)	0.18 (0.17)
9. Moderate - Weak	1.00 (0.35)	0.50 (0.15)
10. Weak - Negligible	0.50 (0.15)	0.18 (0.17)

Note: a_1 – item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension. The *M* and *SD* for '*a*' values are chosen based on Baker's (2001) classification of item discrimination strength (the values are the logistic model parameters values).

[13] Some of the measures with number of items in the neighbourhood of 15 include Geriatric Depression Scale – 15 items version; the Narcissistic Personality Inventory (16 items); the Self-Report Delinquency scale (13 items version); the Mood Disorder Questionnaire (13 items).

Table 2

Item Discrimination and Item Location Parameters for One of the Designed Conditions – Condition 1

Items	a_1	a_2	b_1	b_2
Item 1	1.66	0.25	0.44	0.15
Item 2	2.03	0.35	-0.95	-1.66
Item 3	1.76	0.12	0.08	0.46
Item 4	1.78	0.21	0.19	-1.86
Item 5	1.71	0.14	-0.31	0.81
Item 6	1.96	0.03	-0.18	-0.68
Item 7	1.88	0.23	0.82	-0.89
Item 8	1.78	0.36	0.61	-1.39
Item 9	2.15	0.23	-1.14	-1.49
Item 10	1.92	0.21	1.55	0.28
Item 11	1.7	0.02	-1.08	-0.87
Item 12	1.87	0.09	0.5	1.71
Item 13	2.3	0.19	-1.65	-0.6
Item 14	1.84	0.22	2.01	0.16
Item 15	1.78	0	0.06	-0.92
M (SD)	1.87 (0.18)	0.18 (0.11)	0.06 (1.01)	-0.45 (1.02)

Note. Condition 1 is the condition with a very strong dominant dimension and a negligible secondary dimension. a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension. Item discrimination of the last item was set to 0 on the second dimension in all conditions, as done in multidimensional IRT programs, for estimation reasons (de Ayala, 2009).

3.4 Person Location Parameters Generation

In terms of person location parameter (theta), the thetas for the two dimensions (dominant and secondary dimensions) were generated for 5000 respondents, from the standard normal distribution. Normally distributed traits were of interest in the project, whereas investigating IRT robustness to violated assumptions of unidimensionality/LI, in context of different types of non-normality of the latent traits, is a goal for further stages of this research. Theta parameter was

treated as a random effect, that is, new parameters were generated across the replications/conditions.

Two simulated dimensions were generated as uncorrelated in accordance with the conceptualization of the model. That is, the dominant dimension was conceptualized as representing the target construct and the secondary dimension as representing local dependencies in the measure (i.e., item dependencies remaining after the main latent trait is accounted for). In FA terms, the dominant dimension could be seen as the first factor extracted in initial phase of exploratory factor analysis and the secondary dimension as the second factor extracted, orthogonal to the first common factor by necessity of factor analysis procedure (Gorsuch, 1983; Mulaik, 2009)[14]. As the multidimensional model employed in the current study was not intended as a meaningful multidimensional model (rather as a unidimensional model with a nuisance dimension), factor rotation and positing/estimating correlation between the dimensions (characteristic for 'correlated-traits model') was not of relevance in the current study. The person location parameters and item parameters used in the current study were simulated in R application (R Development Core Team, 2018).

3.5 Multidimensional Model Used for Data Generation

To simulate multidimensional data, two-parameter compensatory MIRT model (Reckase, 2009) was used:

$$P\left(y_{i,j}=1\middle|\theta_i,a_j,d_j\right)=\frac{1}{1+exp[-(a_{j1}\,\theta_{i1}+a_{j2}\theta_{i2}+d_j)]}$$

[14] in an unrotated FA solution.

where the left side of the equation describes the conditional probability that examinee *i*'s response *y* to a dichotomous item *j* is correct, θ_i is latent dimensions vector (with two dimensions θ_{i1} and θ_{i2}), α_j is item slopes vector (with α_{j1} and α_{j2} corresponding to the two latent dimensions) and d_j is the item intercept. The used model is an exploratory MIRT model, in which each item loads on both of the dimensions. Item parameters and theta parameters, for the two dimensions, and in ten designed conditions, were generated as described in the previous section[15]. Compensatory multidimensional model was chosen as more appropriate for the conceptualization of the two dimensions used in the study, compared to noncompensatory model. It assumes that a lower degree of the trait on one dimension can be compensated by higher degree of the trait on the other dimension. Noncompensatory model assumes that for a desired response a certain degree of the trait is required on both dimensions, which is less consistent with the conceptualization of the second dimension in the current study. Item responses were simulated in R application (R Development Core Team, 2018), using 'mirt' R package (Chalmers, 2012).

3.6 Unidimensional IRT Model Estimation

After the sets of item responses were simulated, based on the conditions specified by the study design, unidimensional model estimates were obtained by fitting the unidimensional 2PL IRT model for binary responses, using R package 'ltm' (Rizopoulos, 2006). The 'ltm' package was chosen as it demonstrated good performance in IRT model parameters recovery studies, and it provides estimation of item discrimination/location and person location in R environment. The

[15] The '*d*' parameter (item intercept) was calculated based on the generated '*a*' and '*b*' parameters: $d = -(a_1b_1 + a_2b_2)$, for each of the items.

'ltm' functions fit unidimensional two-parameter logistic model as a special case of a general latent variable model for dichotomous data. The traditional IRT parameters were provided, in logistic metric. Marginal maximum likelihood estimation (MMLE) was utilized for estimation of the item discrimination and location, and for person location estimation, the Bayesian Expected a Posteriori (EAP) method was used (as the most commonly used methods for item parameters and person location estimation, de Ayala, 2009). In terms of number of item parameters, two-parameter model was chosen for the current study. 2PL is a widely used IRT model for binary psychological data (Embretson & Reise, 2000). Furthermore, an important reason for examining 2PL model was due to its comparability to the FA models, traditionally used in psychological science. Specifically, there is a mathematical equivalence between two-parameter normal ogive model and item factor analysis (Kamata & Bauer, 2008; Takane & de Leeuw, 1987).

3.7 Outcome Measures - Distortions in Model Estimates

After the data were simulated, the unidimensional 2PL IRT model was applied to the data and, in each particular experimental condition, the parameter estimates from the model were compared to the corresponding true parameter values, and bias and the root mean square error (RMSE) were calculated. Specifically, bias in estimation of item parameters (i.e., item discrimination and item location) was calculated as a difference between the estimated item parameters and the true parameters (true item discrimination and true item location),

$$bias\ \omega = \frac{\sum_{t=1}^{T} (\hat{\omega} - \omega)}{T}$$

where ω is the parameter under consideration (item discrimination or item location) and T denotes the number of items. Bias in estimation of person location was calculated as a difference between the estimated person location parameter and the true parameter,

$$bias\ \theta = \frac{\sum_{t=1}^{T} \left(\hat{\theta} - \theta\right)}{N}$$

where θ is person location parameter and N denotes the sample size.

RMSE in estimation of item parameters was calculated as

$$RMSE\ \omega = \sqrt{\frac{\sum_{t=1}^{T} (\hat{\omega} - \omega)^2}{T}}$$

where ω is the parameter under consideration (item discrimination or item location parameter) and *T* denotes the number of items. RMSE in estimation of person location was calculated as

$$RMSE\ \theta = \sqrt{\frac{\sum_{t=1}^{N} \left(\hat{\theta} - \theta\right)^2}{N}}$$

where θ is person location parameter and N denotes the sample size. These statistics have been used the most often in evaluation of accuracy and precision of estimated model parameters in IRT simulation studies (Bulut & Sünbül, 2017; Feinberg & Rubright, 2016; Harwell et al., 1996).

In relation to the item parameters estimates, as presented above, bias and RMSE were averaged for the 15 items, and in relation to the person location estimates, bias and RMSE were averaged for 5000 respondents. The two statistics were calculated in 100 replications, in each of the ten conditions, and averaged across the replications in each condition. Additionally, the distortions in the estimates were also investigated at selected different levels of item difficulty and at different theta levels. Furthermore, in order to disentangle the distortions in the model estimates that resulted from the research manipulation according to the study design (i.e., across the ten conditions designed in the study) from the distortions resulting from the IRT estimation method, the distortions in the model estimates were first assessed in several 'strictly unidimensional' conditions. In those conditions, only one dimension was simulated that

corresponded to the dominant dimension in the designed research condition (i.e., very strong, strong, moderate, and weak dimension). The distortions in the unidimensional model estimates obtained in the ten designed conditions were interpreted in comparison to the distortions recorded in the strictly unidimensional conditions[16].

3.8 Eigenvalues

Eigenvalues were calculated from the correlation matrices of the item responses simulated in the ten designed conditions (i.e., at different levels of violations of unidimensionality and LI, as described previously). Analyses were conducted with 100 replications in each of the study conditions. As described in Section 1.2 of this dissertation, eigenvalues were, in the current study, used as indicators of strength of the latent dimensions and their utility as possible "diagnostic indices" in relation to the consequences of the violations of unidimensionality/LI was investigated. Because latent dimensions in the current study have meaning of 'factors' (as the study is conceptualized in common-factor model framework), eigenvalues were calculated based on correlation matrix with communalities on the diagonal. As an initial estimate of communality, the squared multiple correlation was used on the diagonal. Eigenvalues were obtained by using principal-axes factoring extraction method and the analysis was conducted by using 'psych' package in R, as described in Revelle (2017).

The investigation of utility of eigenvalues in previous research was often performed in the context of making decisions in EFA about the number of factors to retain and assessing dimensionality of measures (Ruscio & Roche, 2012; Slocum-Gori & Zumbo, 2011; Velicer,

[16] The same procedure that was used for simulation of the ten experimental conditions (as described in data generation section) was used for simulation of the strictly unidimensional conditions; however, only the first dimension was simulated.

Eaton, & Fava, 2000). When investigated for such a purpose, eigenvalues were traditionally calculated from PCA (based on correlation matrix with total variance on the diagonal, i.e., 1s on diagonal) [17]. Given a different purpose for which eigenvalues were investigated in the current study, and their use as indicators of the strength of the extracted latent dimensions (i.e., factors, not components), obtaining eigenvalues from the correlation matrix with communalities on the diagonal was deemed more appropriate.

Furthermore, because binary item responses were simulated in the current study, eigenvalues were calculated from a tetrachoric correlation matrix[18]. When the utility of eigenvalues was investigated in relation to assessing dimensionality of measures (e.g. number of factors to retain in EFA with categorical data), eigenvalues calculated from a Pearson correlation matrix have been used (Slocum-Gori & Zumbo, 2011) as well as eigenvalues calculated from tetrachoric/polychoric correlation matrix (Cho, Li, & Bandalos, 2009; Weng & Cheng, 2005; Yang & Xia, 2015). The advantage of use of the Pearson correlation matrix is that eigenvalues are easily obtained in any general statistical software, whereas the concern with using the Pearson correlation matrix with categorical data is that the correlations are underestimated (West, Finch & Curran, 1995). Due to convenience and predominant use of Pearson correlation matrix in psychological research, in the current study, four selected indexes were obtained from both tetrachoric and Pearson correlation matrices, and their relations with unidimensional IRT model estimates were compared.

---

[17] The use of eigenvalues from PCA, for purpose of determining number of factors to retain in FA, however, has been questioned (Mulaik, 2009).

[18] FA has been applied to binary/ordered categorical items based on the use of the tetrachoric/polychoric correlation matrix and "latent response variables" (Flora & Curan, 2004; Jöreskog & Moustaki, 2001; Mislevy, 1986; Muthén, 1983; Olsson, 1979; Wirth & Edwards, 2007).

As described previously (in section 1.2 of this dissertation), based on the computed eigenvalues, four eigenvalues-based indexes were calculated (Index 1 – the ratio of first to second eigenvalue index, Index 2 – the difference between first and second eigenvalue index, Index 3 – the proportion of overall standardized variance accounted for by the first common factor, Index 4 – the proportion of overall common variance accounted for by the first common factor), and their relations with the distortions in IRT model estimates were assessed. These four indexes were chosen for the study as they were proposed or studied previously in the context of assessing dimensionality of measures and deciding about the number of factors to retain in EFA, or in context of indexing degree of unidimensionality in measures. For the purpose of indexing degree of unidimensionality, conceptually similar indices, but based on different methods of estimation/different models, were proposed as possibly useful. EVC index (Sijtsma, 2009) corresponds to Index 4 in this study. The percentage of total variance and the percentage of common variance accounted by the general factor, utilized in context of bifactor modeling (Reise et al., 2013, Rodriguez et al., 2016), correspond conceptually to Index 3 and 4 in this study.

In summary, EFA was conducted 100 times in each of the ten study conditions. In each replication, item responses were simulated (as described in Simulated conditions section), correlation matrix of item responses was calculated based on the item responses, EFA was conducted, eigenvalues were obtained, and the four eigenvalue-based indexes formed. Mean and standard deviation were calculated for each of the four indexes across 100 replications. The analyses were first conducted with using Pearson correlation matrix, and then the whole process was repeated with tetrachoric correlation matrix. Correlation matrix with communalities on the diagonal was used in all analyses.

3.9 Data Analysis of the Simulation Results

In terms of the type of data analysis, the focus of the current study was on exploratory data analysis. Bias and RMSE were described/summarized, in tabular and graphical ways, in the ten cells in the study design. The relation between the two factors of interest (i.e., the strength of the dominant dimension and the strength of LD) and bias in the estimation of item discrimination, item location, and person location parameters was explored and summarized – the correlations appropriate for the categorical/ordinal nature of the factors were reported (i.e., Spearman's correlation coefficients in these cases). In relation to eigenvalues and the four eigenvalues-based indexes, they were described/summarized in the ten designed conditions. The relation between the indexes and bias was investigated and reported (i.e., bi-variate Pearson's correlation coefficients, equivalent to standardized regression coefficients in linear regression analysis, were reported). The use of exploratory data analytical approach was deemed appropriate for the research questions asked, and beneficial due to a novel approach employed in the study and the specifics of the research design.

3.10 The Relations Expected Based on the Previous Research

In the context of previous research, it was expected that the closer the designed conditions of unidimensionality/LI violations were to strict unidimensionality, smaller distortions in the IRT model estimates would be, consistent with the previous empirical evidence (Ackerman, 1989; Reckase, 1979; Zhang, 2008) and with the general robustness continuity principle (Huber, 1981; Lind & Zumbo, 1993). In terms of the strength of the dominant dimension, previous research evidence supported the hypothesis that the stronger the dominant dimension was, the smaller the distortions in the unidimensional IRT model estimates would be (Reckase, 1979). In relation to the secondary dimension, based on previous evidence from

research about LD, it is expected that the stronger this dimension was (i.e., stronger LD), the greater the distortions in estimates would be (Chen & Thissen, 1997; Yen, 1993). In various combinations of the strength of the two dimensions, the smallest distortions were expected in the condition characterized by a very strong dominant dimension and a negligible secondary dimension (condition 1), because this condition was seen as closest to strict unidimensionality. Due to the lack of previous evidence, specific hypotheses were not formulated for all combinations of the strength of two dimensions in relation to the distortions in the IRT model estimates – exploring the specifics of these relations and the importance of each of the two factors for the distortions was an important goal embedded in the first purpose of the current study.

In relation to exploring utility of the four eigenvalue-based indexes in predicting distortions in unidimensional model, the expectations were as follows: Index 1 and index 2 are based on the first and second eigenvalues only (Index 1 is a ratio of, whereas Index 2 is a difference in, variances accounted for by the first two factors). By their definition, these two indexes are largest (have the largest value) in the conditions of violations of the two assumptions closest to strict unidimensionality and they were expected to decrease with increase in the amounts of violations (i.e, with decrease in dominant dimension and increase in the secondary dimension). Hence, a negative linear relation was expected in the current study between the size of the two indexes and bias in the unidimensional model estimates (i.e., increase in bias with decrease in the indexes). Index 3 represents the proportion of overall standardized variance accounted for by the first factor and Index 4 represents the proportion of overall common variance accounted for by the first factor. These two indexes, by definition, decrease with decreasing strength of the first, dominant factor. Hence, in terms of the direction of the relations

between these indexes and bias in model estimates, negative linear relations were also expected (increase in bias with decrease in the indexes).

4. Results

4.1 Item Parameters Estimation

After applying the unidimensional IRT model (2PL model) to the item responses in the ten designed conditions, bias and RMSE in the estimation of the item parameters and person parameter were obtained. The results pertaining to the estimation of the item parameters (item discrimination and item location) are presented first.

4.1.1 Item Discrimination

Bias. In terms of bias in item discrimination estimation, in strictly unidimensional conditions (with very strong, strong, moderate, and weak latent dimension), the size of bias ranged from 0.002 to 0.003 (Table 3). In the ten study conditions, with different amounts of unidimensionality and LI violations, the size of bias ranged from 0.007 to 0.38, as presented in Table 4. In all of the conditions, bias was larger than the bias obtained in the corresponding strictly unidimensional conditions, with the smallest bias (i.e., closest to strictly unidimensional conditions) recorded in condition 1 (very strong dominant dimension – negligible secondary dimension) and the largest bias recorded in condition 4 (very strong dominant dimension – strong secondary dimension). In terms of the direction of bias, in all conditions bias was positive, indicating overestimation of the item discrimination parameters[19]. Distribution of bias in the item discrimination estimates, in the ten simulated conditions, is graphically presented in Figure 2.

Regarding the relation between the strength of each of the simulated dimensions and bias, the results pointed towards the greater importance of the strength of the secondary dimension for

[19] In the calculation of bias, true parameters were subtracted from the estimated parameters, with the positive values of bias representing overestimation.

the distortions in the item discrimination estimates, compared to the strength of the dominant dimension. A presence of a strong positive relation between the strength of the secondary dimension and the size of bias was noted; that is, with increase in strength of the secondary dimension, at each level of strength of the dominant dimension, increase in bias in the item discrimination estimates was present, as evident in Table 5 and in Figure 2[20]. Regarding the dominant dimension strength, there was also a certain increase in bias with decrease in the strength of the dominant dimension (a negative relation). As presented in the Table 5 and Figure 2, the changes in bias within rows (across the levels of the dominant dimension strength) were substantially smaller compared to the changes within the columns (across the levels of the secondary dimension strength). The observed trends were confirmed by the calculated correlation coefficients – the magnitude of the relation between the strength of the secondary dimension and bias was 0.90, whereas the magnitude of the negative relation between the strength of the dominant dimension and bias was 0.19[21].

When the ten simulated conditions were ordered according to the amount of bias in the item discrimination parameter (Table 6), the order of conditions reflected the findings regarding the relations between bias and the strength of the dominant and secondary dimensions

---

[20] Different columns in Table 5 and Figure 2 represent different strength of the dominant dimension - within each column increase in strength of the secondary dimension is accompanied by increase in bias.

[21] Semi-partial correlation coefficients were reported, that is, correlations between the strength of one dimension, from which the strength of the other dimension was partialed, and bias. Hence, in the case of the dominant dimension, this correlation represents the percentage of total variation in bias associated with the dominant dimension, with the strength of the secondary dimension partialed from the strength of the dominant dimension (but not partialed from the dependent variable, i.e., bias). This way, in the case of both dominant and secondary dimensions, correlations represent the percentage of *total* variance that each dimension is associated with, making comparison between them feasible. Spearman's correlation coefficients were reported, appropriate for the scale of the factors. $N = 1000$ (100 replications in 10 conditions).

(specifically, the finding that bias is primarily dependent of the secondary dimension strength). The conditions with the largest bias were those with a strong and a moderate secondary dimension (the last three rows in the Table 6). In these conditions, bias could be characterized as large/substantive (e.g., greater than 15% of the average item discrimination[22]). The conditions with a negligible secondary dimension (the first four rows of Table 6) had the smallest amount of bias, which was similar to bias in strictly unidimensional conditions (less than 1% of the average item discrimination).

Finally, bias in the item discrimination estimates was examined at different levels of item location parameter, specifically, at three levels of item location: a) easy items, with item location $\leq$ -0.5, b) medium difficulty items with item location from -0.5 to 0.5, and c) more difficult items, with item location $\geq$ 0.5. In strictly unidimensional conditions, bias was very small at all levels of item location, at different levels of strength of latent dimension (Table 7). In the ten designed conditions, however, a pattern of different bias in item discrimination across different item locations was observed (Table 8). Across the ten conditions, bias was always the smallest in the medium difficulty items, in comparison to the easy and the difficult items. While in most of the conditions these differences were not substantive (i.e., the conclusion about the magnitude of bias would be the same regardless of the item location), in the conditions with the largest bias (condition 4, 3, and 7), the differences were more pronounced. In the two of the conditions, the

[22] '*a*' value of 1 is typically considered an average magnitude of item discrimination (Baker, 2001; Embertson & Reise, 2000). Hence, the size of bias in condition 4 (bias = 0.387) is interpreted as '38.7% of the average item discrimination', whereas in condition 1 (bias = 0.007) bias is interpreted as '0.7% of the average item discrimination'. Interpreting the magnitude of bias in relation to one reference point in all conditions (in this case, in relation to average item discrimination) was deemed more appropriate and meaningful in the current context (as bias in the ten different conditions is compared). Another option is to report relative bias as 'percentage of the true parameter being estimated'. Relative bias reflects the initial size of the parameters, and as such, it was seen as less useful in the context of the current study.

conclusion about bias would differ depending on the level of item location (specifically, bias could be characterized as large in the easy and difficult items, and as small in medium difficulty items). In condition 4, however, with the largest amount of bias, the conclusion about bias would be same regardless of the item location. In the easy and the difficult items, bias was increasing with increase in strength of the secondary dimension, and it was the largest in the conditions with strong and moderate secondary dimension (conditions 4, 3 and 7).

RMSE. Regarding RMSE in estimation of item discrimination parameters, in strictly unidimensional situations with very strong, strong, moderate and weak latent trait, RMSE ranged from 0.049 to 0.078 (Table 3). In the ten study conditions, similar patterns in RMSE were found as in the case of bias. The amount of RMSE was the greatest in the same three conditions in which the bias was the largest: conditions with strong secondary dimension (condition 4) and with moderate secondary dimension (condition 3 and 7), ranging from 0.278 to 0.493 (Table 4). RMSE was the smallest in the conditions with negligible secondary dimension (condition 1, 5, 8 and 10), ranging from 0.055 to 0.078, and in these conditions, RMSE was very similar to RMSE in the strictly unidimensional conditions (Table 3). In the ten simulated conditions, the amount of RMSE was increasing with increase in strength of the secondary dimension, within different levels of the dominant dimension. In terms of ordering of the ten conditions by the size of RMSE, except slight differences in regard to the conditions with negligible secondary dimensions (with very small amount of RMSE recorded), the other conditions were ordered in the same way according to size of RMSE as they were according to size of bias in item discrimination estimates (Table 6).

Table 3

Bias and RMSE in Estimation of Item Discrimination and Item Location in Strictly Unidimensional Conditions

Latent Trait Strength	**Item Discrimination**		**Item Location**	
	Bias *M (SD)*	***RMSE*** *M (SD)*	***Bias*** *M (SD)*	***RMSE*** *M (SD)*
Very strong	0.003 (0.02)	0.078 (0.02)	<0.001 (0.02)	0.032 (0.01)
Strong	0.003 (0.02)	0.064 (0.01)	<0.001 (0.02)	0.041 (0.01)
Moderate	0.003 (0.02)	0.053 (0.01)	-0.001 (0.02)	0.056 (0.01)
Weak	0.002 (0.01)	0.049 (0.01)	0.002 (0.03)	0.112 (0.03)

Note: In strictly unidimensional conditions, only one latent trait was simulated that corresponded to the strength of the dominant traits simulated in the ten experimental conditions (no secondary dimension was simulated). *M* and *SD* are across 100 replications in each condition.

Table 4

Bias and RMSE in Estimation of Item Discrimination and Item Location in the Ten Designed Conditions

Designed Conditions	Item Discrimination		Item Location	
	Bias *M (SD)*	**RMSE** *M (SD)*	**Bias** *M (SD)*	**RMSE** *M (SD)*
1. Very strong - Negligible	0.007 (0.02)	0.078 (0.02)	-0.058 (0.01)	0.134 (0.01)
2. Very strong - Weak	0.053 (0.02)	0.109 (0.02)	-0.102 (0.01)	0.274 (0.01)
3. Very strong - Moderate	0.187 (0.02)	0.278 (0.02)	-0.220 (0.01)	0.489 (0.01)
4. Very strong - Strong	0.384 (0.02)	0.493 (0.01)	-0.245 (0.01)	0.613 (0.01)
5. Strong - Negligible	0.009 (0.01)	0.059 (0.01)	-0.085 (0.01)	0.183 (0.01)
6. Strong - Weak	0.072 (0.02)	0.114 (0.02)	-0.139 (0.01)	0.349 (0.01)
7. Strong - Moderate	0.250 (0.02)	0.341 (0.02)	-0.275 (0.01)	0.593 (0.01)
8. Moderate - Negligible	0.015 (0.01)	0.055 (0.01)	-0.095 (0.01)	0.232 (0.01)
9. Moderate - Weak	0.100 (0.01)	0.144 (0.01)	-0.145 (0.01)	0.477 (0.02)
10. Weak - Negligible	0.027 (0.01)	0.065 (0.01)	-0.216 (0.03)	0.449 (0.03)

Note. *M* and *SD* are across 100 replications in each condition.

Table 5

Bias in Item Discrimination Estimates by the Strength of Dominant and Secondary Dimension

		Dominant Dimension Strength			
		Very Strong	*Strong*	*Moderate*	*Weak*
Secondary Dimension Strength	*Negligible*	0.007	0.009	0.015	0.027
	Weak	0.053	0.072	0.100	
	Moderate	0.187	0.250		
	Strong	0.384			

Note. The values are the mean bias in the ten created conditions (over 100 replications).

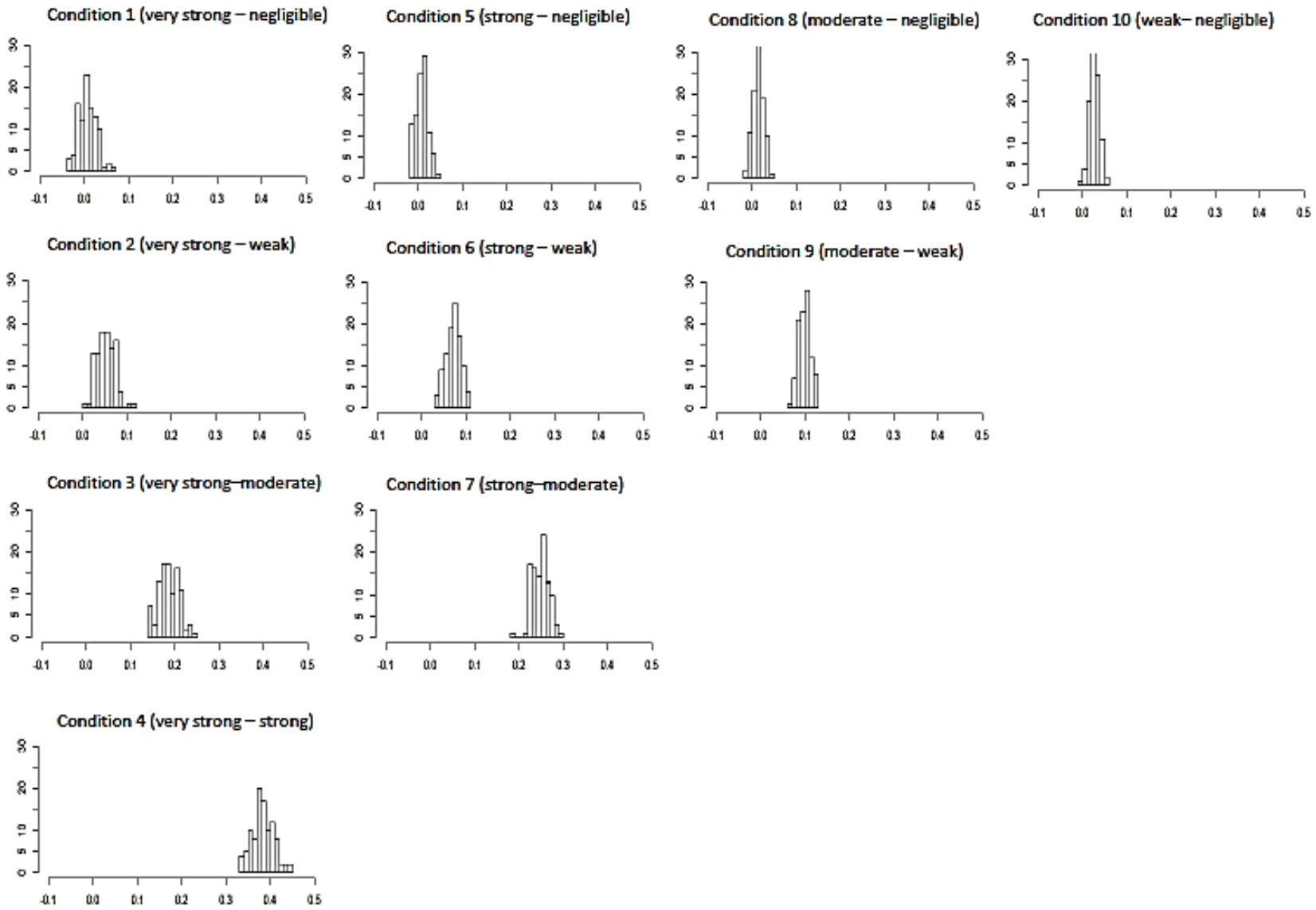


Figure 2. Distribution of bias in the item discrimination parameter estimates in the ten designed condition (across 100 replications). Bias is on the x-axis and frequency is on the y-axis. Individual graphs are also included in Appendix B.

Table 6

Ten Designed Conditions Ordered by the Size of Bias in Item Discrimination Parameter

Designed Conditions	Item Discrimination		Item Location	
	Bias	**RMSE**	**Bias**	**RMSE**
1. Very strong - Negligible	0.007	0.078	-0.058	0.134
5. Strong - Negligible	0.009	0.059	-0.085	0.183
8. Moderate - Negligible	0.015	0.055	-0.095	0.232
10. Weak - Negligible	0.027	0.065	-0.216	0.449
2. Very strong - Weak	0.053	0.109	-0.102	0.274
6. Strong - Weak	0.072	0.114	-0.139	0.349
9. Moderate - Weak	0.100	0.144	-0.145	0.477
3. Very strong - Moderate	0.187	0.278	-0.220	0.489
7. Strong - Moderate	0.250	0.341	-0.275	0.593
4. Very strong - Strong	0.384	0.493	-0.245	0.613

Note. The values are the mean bias and RMSE in the ten conditions (over 100 replications).

Table 7

Bias in Estimation of Item Discrimination and Item Location across Different Levels of Item Location in Strictly Unidimensional Conditions

Designed Conditions	**Bias in Item Discrimination**			**Bias in Item Location**		
	Easy Items	Medium Items	Difficult Items	Easy Items	Medium Items	Difficult Items
Very strong	<0.001	0.002	0.006	<0.001	-0.001	-0.001
Strong	0.003	0.004	0.003	0.002	-0.001	-0.001
Moderate	0.003	0.002	0.004	-0.001	0.001	-0.002
Weak	0.003	0.001	0.001	-0.002	-0.001	0.010

Note. Easy items: $b = \leq -0.5$; medium difficulty items: $-0.5 < b < 0.5$; difficult items: $b \geq 0.5$ (b – item location). The values are the mean bias across 100 replications in each condition.

Table 8

Bias in Estimation of Item Discrimination and Item Location across Different Levels of Item Location in the Ten Designed Conditions

Designed Conditions	Bias in Item Discrimination			Bias in Item Location		
	Easy Items	Medium Items	Difficult Items	Easy Items	Medium Items	Difficult Items
1. Very strong - Negligible	0.016	-0.001	0.012	-0.082	-0.034	-0.056
2. Very strong - Weak	0.072	0.025	0.070	-0.159	-0.077	-0.066
3. Very strong - Moderate	0.242	0.054	0.266	-0.220	-0.184	-0.258
4. Very strong - Strong	0.386	0.277	0.488	-0.230	-0.259	-0.247
5. Strong - Negligible	0.013	0.000	0.016	-0.145	-0.045	-0.072
6. Strong - Weak	0.090	0.031	0.090	-0.232	-0.102	-0.083
7. Strong - Moderate	0.333	0.069	0.346	-0.268	-0.219	-0.339
8. Moderate - Negligible	0.018	0.004	0.022	-0.156	-0.065	-0.064
9. Moderate - Weak	0.121	0.066	0.113	-0.250	-0.155	-0.031
10. Weak - Negligible	0.030	0.009	0.043	-0.270	-0.166	-0.212

Note. Easy items: $b = \leq -0.5$; Medium Difficulty Items: $-0.5 < b < 0.5$; Difficult Items: $b \geq 0.5$.
b - Item location. The values are the mean bias over 100 replications in each condition.

4.1.2 Item Location

Bias. The obtained results pertaining to bias in the item location estimation follow a similar pattern as the results obtained regarding bias in item discrimination estimation, with some differences noted. In all conditions of violations of the assumptions, bias in the item location estimates was negative, indicating underestimation of item location parameter, for a difference from overestimation in the case of bias in item discrimination (Table 4). Graphical presentations of the distributions of bias in the ten conditions are included in Figure 3.

In strictly unidimensional conditions, bias in the item location estimates was very small (the range of bias in was from -0.001 to 0.002; Table 3). In the ten conditions of violations of the assumptions, the amount of bias was increasing with increase in strength of the secondary dimension, at each level of strength of the dominant dimension (Table 4). The conditions with the largest amount of bias were those with strong secondary dimension (condition 4) and with moderate secondary dimension (condition 3 and 7). In the three conditions with the largest bias, somewhat different order of the conditions was obtained in comparison to the order according to bias in item discrimination estimation. Furthermore, different from item discrimination estimation, substantial bias in the item location parameter estimates was also recorded in condition 10 (weak – negligible). These differences were reflected in the magnitude of the correlations between bias in the item location estimates and strength of the two dimensions – greater magnitude of the correlation was recorded with strength of the dominant dimension (correlation coefficient of 0.54) compared to the magnitude of the correlation obtained in the case of item discrimination[23]. The relation was negative, as was in the case of item

[23] Spearman's semi-partial correlation coefficients were reported (N = 1000).

discrimination. The correlation coefficient of 0.91 was recorded with the strength of the secondary dimension, very similar as in the case of item discrimination. The order of the ten conditions according to bias in item location was somewhat different from the order of the conditions according to bias in item discrimination (Table 6). Finally, the results about bias in the item location estimates, at different levels of item location, are presented in Table 7 and Table 8. A pattern of greater bias in the easier and the more difficult items than in the medium difficulty items, obtained in relation to bias in the item discrimination, was not as clear as in bias in the item location. In some of the conditions, with larger amount of bias in the item location estimates, the pattern was noticeable (condition 10, 3, and 7; Table 8).

RMSE. In relation to RMSE in estimation of the item location parameters, the same pattern of the results was obtained as in relation to bias in estimation of the item location parameters. The range of RMSE in strictly unidimensional conditions was from 0.032 to 0.112 (Table 3). In the ten designed conditions, the largest amount of RMSE was recorded in the conditions with strong secondary dimension (condition 4) and with moderate secondary dimension (condition 3 and 7) - RMSE in these three conditions was 0.613, 0.593, 0.489, respectively (Table 4). The order of the ten conditions based on RMSE in item location estimation and based on bias in item location estimation was very similar (Table 6).

Table 9

Bias in Item Location Estimates by Strength of Dominant and Secondary Dimension

		Dominant Dimension Strength			
		Very Strong	Strong	Moderate	Weak
Secondary Dimension Strength	Negligible	-0.058	-0.085	-0.095	-0.216
	Weak	-0.102	-0.139	-0.145	
	Moderate	-0.220	-0.275		
	Strong	-0.245			

Note. The values are the mean bias in the ten created conditions, across 100 replications.

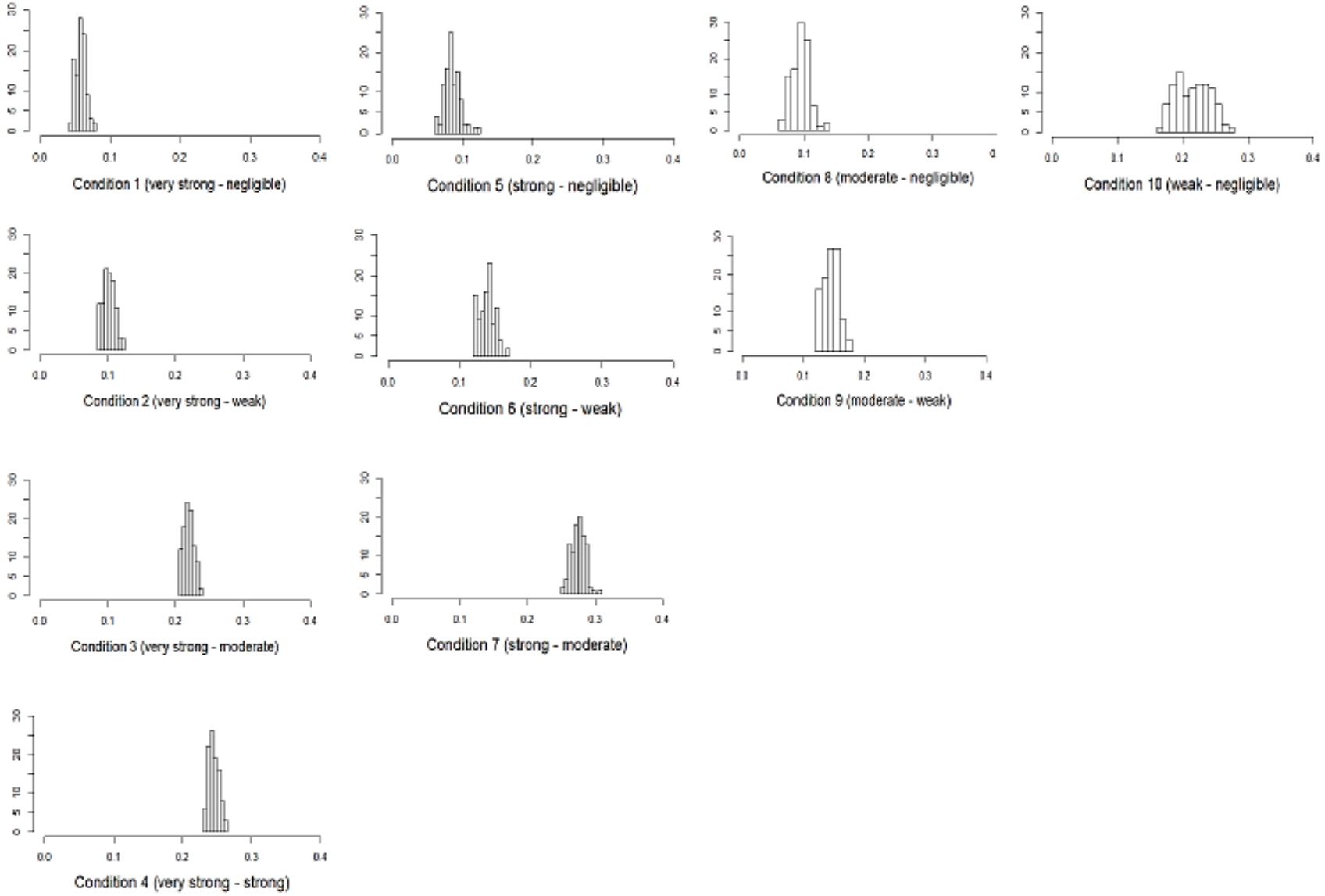


Figure 3. Distribution of bias in the item location parameter estimates in the ten designed condition (across 100 replications). Bias is on the x-axis and frequency is on the y-axis. *Note*: absolute values are plotted in this distribution graph for simplicity of presentation (values of bias are negative). Individual graphs are also included in Appendix B.

4.2 Person Location Parameter Estimation

In relation to person location estimation, in strictly unidimensional conditions bias in the estimates was very close to zero, as in case of bias in the item parameters estimates (Table 10). However, different from the item parameters estimation, no bias was found in the person location estimates in the ten designed conditions. The average bias was very close to zero in each of the conditions, suggesting that there was no systematic distortion in the person location estimates (Table 11). However, in terms of the absolute size of distortions, as indicated by RMSE, distortions were recorded in strictly unidimensional conditions (RMSE from 0.377 to 0.725; Table 10), as well as in each of the ten designed conditions (a range of RMSE from 0.386 to 0.747; Table 11).

Further examination, at different levels of latent trait, shed more light on the above reported results. Bias was summarized at three levels of latent trait: a) low range of trait, with theta $\leq$ -1, b) moderate range of latent trait, with theta from -1 to 1, and c) higher range of latent trait, with theta $\geq$ 1. The results revealed that the findings about no bias applied only to the moderate range of latent trait, in all of the conditions. At the lower range of latent trait, however, as well as at the higher range of latent trait, there was systematic error present in the latent trait estimates - in strictly unidimensional conditions (Table 12) as well as in each of the ten designed conditions (Table 13). The results consistently indicated *overestimation* of the latent trait at the lower range of the trait, and *underestimation* at the higher range of the trait. There was increase in bias in the person location estimates, at both lower and higher ranges of latent trait, with increase in strength of the secondary dimension; however, the trend was stronger in relation to the strength of the dominant dimension. Specifically, bias increased with decrease in strength of the dominant dimension (negative relation), with the greatest amount of bias recorded in

condition 10 (with a weak dominant dimension): bias = 0.803 at the lower level of the latent trait and bias = -0.821 at the higher level of the latent trait. A very similar amount of bias was recorded in the strictly unidimensional condition with a weak dominant dimension (bias = 0.798 at the lower range of the latent trait and bias = -0.797 at the higher range of the latent trait; Table 12). The magnitude of the correlations between strength of the two dimensions and bias in the person location estimation (at the low and the high ranges of latent trait) supported greater importance of the strength of the dominant dimension for bias in the person location estimation. The magnitude of the correlation between the strength of the dominant dimension and bias at the low level of latent trait was 0.97, and at the high level of latent trait the magnitude of the correlation was 0.93. The magnitude of the correlation between the strength of the secondary dimension and bias at the low level of latent trait was 0.59 and at the high level of latent trait, the magnitude of the correlation was 0.69[24].

[24] Spearman's semi-partial correlations were reported (N = 1000).

Table 10

Bias and RMSE in Estimation of Person Location Parameter in Strictly Unidimensional Conditions

Latent Trait Strength	**Bias** *M* (*SD*)	**RMSE** *M* (*SD*)
Very strong	0.001 (0.01)	0.377 (0.003)
Strong	0.001 (0.01)	0.430 (0.004)
Moderate	0.001 (0.01)	0.524 (0.010)
Weak	0.001 (0.01)	0.725 (0.010)

Note: In strictly unidimensional conditions, no secondary dimension was simulated. *M* and *SD* are across 100 replications in each condition.

Table 11

Bias and RMSE in Estimation of Person Location Parameter in the Ten Designed Conditions

Designed Conditions	**Bias** *M* (*SD*)	**RMSE** *M* (*SD*)
1. Very strong - Negligible	<0.001 (0.001)	0.386 (0.004)
2. Very strong - Weak	<0.001 (0.001)	0.439 (0.004)
3. Very strong - Moderate	<0.001 (0.001)	0.560 (0.01)
4. Very strong - Strong	<0.001 (0.002)	0.679 (0.004)
5. Strong - Negligible	<0.001 (<0.001)	0.445 (0.004)
6. Strong - Weak	<0.001 (<0.001)	0.514 (0.004)
7. Strong - Moderate	<0.001 (0.001)	0.658 (0.01)
8. Moderate - Negligible	<0.001 (<0.001)	0.541 (0.01)
9. Moderate - Weak	<0.001 (<0.001)	0.626 (0.01)
10. Weak - Negligible	<0.001 (<0.001)	0.747 (0.01)

Note. *M* and *SD* in each condition are across 100 replications.

Table 12

Bias in Estimation of Person Location Parameter across Different Levels of Person Location in Strictly Unidimensional Conditions

Designed Conditions	**Bias**		
	Low θ	Medium θ	High θ
Very strong	0.207	0.003	-0.218
Strong	0.300	-0.005	-0.276
Moderate	0.433	-0.006	-0.404
Weak	0.798	0.000	-0.797

Note. Low level of latent trait: $\theta \leq -1$; medium level: $-1 < \theta < 1$; high level: $\theta \geq 1$. The values are the mean bias across 100 replications in each condition.

Table 13

Bias in Estimation of Person Location across Different Levels of Person Location in the Ten Designed Conditions

Designed Conditions	Bias		
	Low θ	Medium θ	High θ
1. Very strong - Negligible	0.211	0.004	-0.230
2. Very strong - Weak	0.242	0.005	-0.262
3. Very strong - Moderate	0.323	0.008	-0.358
4. Very strong - Strong	0.421	0.011	-0.466
5. Strong - Negligible	0.301	-0.002	-0.292
6. Strong - Weak	0.345	-0.001	-0.340
7. Strong - Moderate	0.454	0.002	-0.461
8. Moderate - Negligible	0.435	-0.001	-0.427
9. Moderate - Weak	0.494	0.000	-0.493
10. Weak - Negligible	0.803	0.004	-0.821

Note. Low level of latent trait: $\theta \leq -1$; medium level: $-1 < \theta < 1$; high level: $\theta \geq 1$. The values are the mean bias across 100 replications in each condition.

4.3 Eigenvalue-Based Indexes and Distortion in Model Estimates

Using eigenvalues from principal-axes factoring of the item responses' correlation matrices in the ten simulated conditions, the four selected indexes were calculated (Table 14). The indexes based on tetrachoric and the indexes based on Pearson correlations differed in the same conditions, with the size of the indexes based on the tetrachoric correlation matrix larger compared to the size of the indexes based on the Pearson correlation matrix. Regardless of the type of correlation matrix used in the analyses, some unexpected patterns of increase/decrease in the size of certain indexes were obtained in the designed conditions. For example, in relation to Index 1 (the ratio of first to second eigenvalue), the index was decreasing with increase in strength of the secondary dimension, as expected by the definition of the index, at all levels of strength of the dominant dimension, (e.g., conditions 1 – 4, Table 14); however, decrease in the index was not recorded with decrease in strength of the dominant dimension (e.g., conditions 1, 5, 8, 10). A counterintuitive pattern was therefore noted, in which the ratio of first to second eigenvalue was greater in condition 5 than in condition 1. Therefore, further exploration of the observed patterns was performed, by investigating the obtained eigenvalues in greater detail (Table 15), before proceeding with the investigation of the main issue concerning the relations between the eigenvalue-based indexes and the IRT model estimates.

In relation to the size of the first and second eigenvalues in the simulated conditions (Table 15), some trends should be highlighted. As expected, the first eigenvalue was increasing with increase in strength of the dominant dimension (e.g., conditions 1, 5, 8 and 10), and in regard to the second eigenvalue, it was increasing with increase in strength of the secondary dimension (e.g., conditions 1 to condition 4). However, the recorded patterns also suggested that the size of the eigenvalues, both the first and the second, was influenced by the strength of the

both dimensions. For example, the first eigenvalue was increasing with increase in strength of the secondary dimension and the second eigenvalue was increasing with increase in strength of the dominant dimension. Such trends, likely related to the type of correlation matrix used in the analyses[25], had implications on the patterns of the calculated indexes, with the greatest impact on Index 1. This index is a 'ratio'; hence, its size by definition depended on the estimates of the strength of *both* dimensions to a greater degree compared to the other calculated indexes[26]. The indexes primarily reflective of strength of the dominant dimension, overall, demonstrated the expected patterns in the designed condition, systematically decreasing with decrease in strength of the dominant dimension.

The results regarding the relations between the four eigenvalues-based indexes and bias in IRT model estimates are summarized in Table 16. In regard to the item parameters estimation, the indexes demonstrated weak to moderate relations with bias in the item parameters estimates. In regard to bias in person location estimates, three of the indexes demonstrated strong relations with bias in the estimates of person location parameter; specifically, Index 2, Index 3, and Index 4. In cases when strong relations were recorded between the indexes and bias, the direction of the relations was in the expected direction, that is, negative correlation was recorded indicating increase in bias with decrease in the indexes.

[25] Correlation matrix with communalities on the diagonal was used instead of correlation matrix with total variance (e.g., 1) on the diagonal. Overall common variance changes across the simulated conditions while the overall variance does not.

[26] Index 2 is also, by definition, reflective of strength of both dimensions; however, it is a 'difference' index, and as such, its size was primarily determined by the size of the stronger dimension (i.e., first dimension).

Table 14

Eigenvalues-Based Indexes Based on Pearson and Tetrachoric Correlation Matrices

Simulated Conditions	Based on Pearson Correlations				Based on Tetrachoric Correlations			
	Index 1 *M* (*SD*)	Index 2 *M* (*SD*)	Index 3 *M* (*SD*)	Index 4 *M* (*SD*)	Index 1 *M* (*SD*)	Index 2 *M* (*SD*)	Index 3 *M* (*SD*)	Index 4 *M* (*SD*)
1. Very strong - Negligible	7.66 (0.32)	3.72 (0.04)	0.29 (0.003)	0.81 (0.01)	27.13 (9.76)	7.56 (0.21)	0.53 (0.01)	0.88 (0.03)
2. Very strong - Weak	6.57 (0.22)	3.58 (0.04)	0.28 (0.003)	0.79 (0.01)	21.38 (10.26)	7.50 (0.40)	0.54 (0.01)	0.86 (0.04)
3. Very strong - Moderate	5.54 (0.21)	3.51 (0.04)	0.29 (0.003)	0.75 (0.01)	14.40 (8.5)	7.65 (0.47)	0.57 (0.01)	0.84 (0.04)
4. Very strong - Strong	4.48 (0.14)	3.41 (0.04)	0.29 (0.003)	0.73 (0.01)	12.84 (6.56)	8.13 (0.44)	0.60 (0.01)	0.81 (0.04)
5. Strong - Negligible	15.19 (1.17)	3.21 (0.04)	0.23 (0.003)	0.84 (0.01)	28.32 (9.85)	5.72 (0.17)	0.40 (0.01)	0.87 (0.02)
6. Strong - Weak	11.43 (0.83)	3.17 (0.04)	0.23 (0.003)	0.83 (0.01)	26.29 (8.25)	5.95 (0.15)	0.42 (0.01)	0.86 (0.02)
7. Strong - Moderate	8.13 (0.40)	3.25 (0.04)	0.25 (0.003)	0.79 (0.01)	24.80 (3.12)	6.65 (0.10)	0.46 (0.01)	0.86 (0.01)
8. Moderate - Negligible	16.44 (2.12)	2.11 (0.04)	0.15 (0.003)	0.79 (0.01)	19.55 (5.76)	3.69 (0.14)	0.26 (0.01)	0.80 (0.02)
9. Moderate - Weak	13.81 (1.58)	2.24 (0.04)	0.16 (0.003)	0.79 (0.01)	19.81 (4.45)	4.1 (0.11)	0.29 (0.01)	0.81 (0.02)
10. Weak - Negligible	8.32 (1.66)	0.81 (0.04)	0.06 (0.002)	0.61 (0.02)	8.06 (1.82)	1.31 (0.04)	0.10 (0.004)	0.61 (0.03)

Note. Index 1 – the ratio of first to second eigenvalue index; Index 2 – the difference between first and second eigenvalue index; Index 3 - the proportion of overall standardized variance explained by first common factor index; Index 4 - the proportion of common variance explained by first common factor index. *M* and *SD* for each index, in the ten conditions, are across 100 replications.

Table 15

First and Second Eigenvalues Based on Pearson and Tetrachoric Correlation Matrices

Simulated Conditions	Based on Pearson Correlations		Based on Tetrachoric Correlations	
	1st Eigenvalue *M (SD)*	2nd Eigenvalue *M (SD)*	1st Eigenvalue *M (SD)*	2nd Eigenvalue *M (SD)*
1. Very strong - Negligible	4.28 (0.04)	0.56 (0.02)	7.92 (0.08)	0.36 (0.22)
2. Very strong - Weak	4.23 (0.04)	0.65 (0.02)	8.07 (0.10)	0.57 (0.47)
3. Very strong - Moderate	4.29 (0.04)	0.77 (0.03)	8.51 (0.11)	0.86 (0.53)
4. Very strong - Strong	4.39 (0.04)	0.98 (0.03)	9.05 (0.10)	0.92 (0.49)
5. Strong - Negligible	3.44 (0.04)	0.23 (0.02)	5.99 (0.07)	0.26 (0.17)
6. Strong - Weak	3.47 (0.04)	0.31 (0.02)	6.23 (0.08)	0.28 (0.15)
7. Strong - Moderate	3.70 (0.04)	0.46 (0.02)	6.93 (0.08)	0.28 (0.04)
8. Moderate - Negligible	2.25 (0.04)	0.14 (0.02)	3.93 (0.07)	0.24 (0.14)
9. Moderate - Weak	2.42 (0.04)	0.18 (0.02)	4.34 (0.07)	0.24 (0.09)
10. Weak - Negligible	0.92 (0.03)	0.12 (0.04)	1.52 (0.06)	0.21 (0.09)

Note. *M* and *SD* are across 100 replications in each condition.

Table 16

The Magnitude of the Relations between Four Eigenvalues-Based Indexes (Based on Pearson and Tetrachoric Correlation Matrices) and Bias in the Unidimensional IRT Model Estimates

	Indexes Based on Pearson Correlation Matrix				**Indexes Based on Tetrachoric Correlation Matrix**			
	Index 1	Index 2	Index 3	Index 4	Index 1	Index 2	Index 3	Index 4
Item Discrimination Bias	-0.56	0.32	0.44	-0.18	-0.22	0.49	0.53	0.09
Item Location Bias	-0.51	-0.14	-0.03	-0.54	-0.37	0.03	0.07	-0.32
Person Location Bias[a]	0.05	-0.89	-0.82	-0.82	0.47	-0.78	-0.75	-0.87

Note: Bi-variate correlation coefficients between the indexes and bias are reported (Pearson's *r*). Values based on *N* = 1000 (100 replications in 10 conditions). Index 1 – the ratio of first to second eigenvalue index; Index 2 – the difference between first and second eigenvalue index; Index 3 – the proportion of overall standardized variance explained by the first common factor index; Index 4 – the proportion of common variance accounted by the first common factor index.

[a] In case of person location bias, reported coefficient is the average of the correlations between the index and bias at the lower end and at the higher end of latent trait.

5. Discussion

Responses to items in psychological measures are multiply determined; therefore, certain violations of the strict assumptions of unidimensionality and local independence (LI) are common in the measures intended to measure a single construct (McDonald, 1999; Nandakumar, 1991; Stout, 1987; Traub, 1983). The current study was conducted with a general purpose of providing more information about the violations of the two assumptions and the accuracy of the models that assume unidimensionality and LI. The violations of the assumptions were studied in a unifying approach, in which violations of LI were modeled in a dimensionality framework (Ip, 2010; Lazarsfeld, 1950). A secondary, nuisance latent dimension was simulated in addition to the dominant dimension underlying item responses, representing item dependencies not accounted by the dominant dimension. The strength of the two dimensions was then systematically varied, creating ten conditions that depicted different levels of violations of unidimensionality and LI. The main goal of the study was to explore, in greater detail, the relation between different conditions of unidimensionality and LI violations and the distortions in the unidimensional IRT model estimates (specifically, bias and RMSE in the 2PL IRT estimation of item discrimination, item location, and person location). Furthermore, eigenvalues from EFA (i.e., four eigenvalues-based indexes) were utilized in the study as indexes of the amounts of violations of the two assumptions, and their utility was assessed in formulating rules/recommendations concerning the use of unidimensional models when violations are present in measures.

The main research question addressed in the study was informed by previous research about the consequences of the violations of unidimensionality/strict unidimensionality for unidimensional IRT model estimates, predominantly conducted in the educational/achievement assessment context (Ackerman, 1989; Crişan, et al. 2017; Drasgow & Parsons, 1983; Ip, 2010;

Kahraman, 2013; Kirisci et al., 2001; Reckase, 1979; Sahin et al, 2015; Zhang, 2008). A large body of research has been conducted on the topic; however, the results of the research have been characterized as "far from conclusive" (Kahraman, 2013). The need for further details about the robustness of unidimensional IRT models has been highlighted (Ip, 2010). Based on summarizing the findings of the studies and identifying the most recent research trends, methodological decisions were made in the current study that were deemed possibly useful in further development of such trends and in providing new evidence for the field. Therefore, the decisions were made to focus specifically on the forms of violations more typical of unidimensional measures (i.e., weaker forms of violations; Ip, 2010; Kahraman, 2013; Zhang, 2008), to focus on complex items and within-item multidimensionality (Kahraman, 2013; Reise, et al., 2013b; Sahin et al, 2015; Zhang, 2008), and to adopt a unifying framework for studying violations of unidimensionality and LI (Ip, 2010).

Several different conceptual and methodological decisions were made in this study compared to previous unidimensional IRT robustness research. A more general model was utilized, in which LD was conceptualized as residuals (i.e., as a nuisance dimension uncorrelated to the dominant dimension), as opposed to the predominantly used correlated-traits model. The conditions of the violations of unidimensionality/LI were created based on varying the strength of the simulated latent dimensions (in a specific partially-crossed design). Consequently, in investigating the ways of providing recommendations about the distortions in parameters estimates (at different levels of violations), the utility of the indices of the strength of the latent dimensions was investigated (i.e., eigenvalues-based indices), as opposed to forming recommendations based on the size of the correlation between the latent dimension. Out of various, possibly useful, indices of the strength of the latent dimensions (Reise et al., 2013a;

Rodriguez et al., 2016), eigenvalues from EFA were chosen because of their convenience and their common use in applied research. Such indices were traditionally investigated in the context of assessing dimensionality/deciding how many factors to retain in FA solution (Slocum-Gori & Zumbo, 2011; Yang & Xia, 2015; Zopluoglu & Davenport, 2017). The above described methodological decisions were made in the current study as possibly useful in contributing more evidence toward a) formulating general rules about relations between the levels of violations and the size of the distortions in unidimensional model estimates and b) finding more pragmatic and convenient ways to express and communicate such rules to applied researchers. The conditions relevant for general psychological measures were given special consideration – the research was conducted with a shorter measure (15 items), items of greater "complexity" were simulated, and the sources of LD were not specified.

5.1 Item Parameters Estimation

A strong relation was found between the amounts of violations – as defined by the strength of LD, and the distortions in the item parameters estimation (e.g., a correlation of 0.90 was recorded between the strength of LD and bias in the item discrimination estimates, and a correlation of 0.91 between the strength of LD and bias in the item location estimates, after controlling for the strength of dominant dimension). The strong relation, found between the strength of LD and the distortions in the item parameter estimation, is consistent with the findings of the line of research about LD, in which specific types of LD were studied (e.g., testlets) and LD was measured in different ways (DeMars, 2012; Steinberg & Thissen, 1996; Yen, 1993; Wainer & Thissen, 1996). Negative consequences of LD, for the use of IRT models, were highlighted in this research.

In the current study, in all of the simulated conditions, item discrimination parameters were overestimated, with overestimation systematically increasing with increase in strength of LD. The strength of the dominant dimension had less importance for the accuracy of the item parameter estimates, compared to the strength of LD. Due to overestimation of item discrimination, the resulting item characteristic curves (ICC's) are biased (i.e., steeper compared with the true ICCs), and item- and test-precision are overestimated. Overestimation of reliability and test information functions, as a result of different specific types of LD in measures, was indicated in LD literature (Sireci et al., 1991; Wainer et al., 2007; Wainer & Thissen, 1996; Yen, 1993; Zenisky, Hambleton, & Sireci, 2002). In some studies focused on violation of unidimensionality (Kahraman, 2013; Sahin, et al, 2015), overestimation of item discrimination was also found as a consequence of increase in multidimensionality (as defined in the particular context). The investigation conducted in the current study, at different levels of item difficulty, revealed that overestimation of the item discrimination estimates was consistently larger for the easier and more difficult items. Therefore, increase in the ICC steepness with increase in strength of LD may be largest in the easier and more difficult items, and the consequences for item information functions and related uses of IRT, may be greatest in such items.

In terms of the magnitude of bias in the item parameter estimates, in most of the simulated conditions, bias could be characterized as "not substantial". In a range of violations that typically are found in well-designed psychological measures intended to measure one construct, the result of the study supported characterizing bias in item parameter estimates as not substantial. In the designed conditions characterized by the presence of a negligible or a weak LD, item discrimination would be misestimated by less than 0.15 points (e.g., less than 15% of the average item discrimination value). In the conditions with a weak/negligible LD, average

item discrimination of 0.5 and less was simulated on the secondary dimensions. In the conditions with a strong and a moderate LD, bias exceeded 0.15 points in item discrimination (average item discrimination of 0.9 and larger, was simulated on the secondary dimensions in the conditions with a moderate/strong LD). In relation to the item location estimation, item location was misestimated for less than 0.2 SD (0.2 points on a standard normal scale) in most of the conditions. In the conditions with a 'moderate' and a 'strong' nuisance dimension, bias increased, but without exceeding 0.3 SD points in any of the conditions. It should be highlighted that, in the context of the real measures, the answer to the question whether bias is "small" or "large" is tied to the practical consequences that may result from the use of the distorted estimates. That is, for different practical consequences, different amount of distortions in IRT model estimates may be relevant (Zhao & Hambleton, 2017). Along those lines, the data about bias provided in the current study should be interpreted with such specific practical contexts in mind.

5.2 Person Location Estimation

In relation to the person location estimation, some important differences in the obtained results were noted compared to the results obtained about item parameters estimation. The analyses based on the whole range of the target latent trait suggested that there was no bias in the person location estimates in any of the conditions. More specific examination that followed (i.e., summarizing bias at different levels of the latent trait), revealed that in all conditions of assumptions violations, the finding of no bias held only for the 'average' level of the latent trait that we intended to measure. Bias was present at more extreme levels of person location, that is, at the 'low' and 'high' levels of latent trait. The person location was systematically overestimated at the lower end of the latent trait, and underestimated at the higher end of latent

trait. In the IRT robustness studies in which bias was not examined across the levels of latent trait, it has been often reported that person location estimates were free of bias but negative effects on the precision of estimates were reported (Crişan, et al. 2017). The other studies reported negative effects of violations of unidimensionality/LI on the accuracy in the estimation of person location and that error in person location estimation decreased as correlation between the latent dimensions increased (Sahin, et al, 2015; Zenisky et al, 2002; Zhang, 2009).

For a difference from the item parameter estimation, bias in person location was present in strictly unidimensional conditions too (i.e., the same pattern of differential bias was revealed). In the conditions of assumptions violations, as well as in unidimensional conditions, bias was systematically increasing with decrease of the strength of the dominant dimension (negative correlation of 0.97 and 0.93 between bias and the strength of the dominant dimension at the low and high levels of latent trait, respectively). The condition with a weak dominant dimension was most problematic in terms of accuracy of the person location estimates. The strength of the dominant factor was in IRT literature suggested as the most important for accuracy of IRT estimates (Embertson & Reise, 2000), and the indexes based on the strength of the dominant factor have been proposed as indicators of appropriateness of use of unidimensional models, such as the proportion of overall variance explained by the first component in PCA (Reckase, 1979); the proportion of overall variance explained by general factor in bifactor models (Reise et al., 2013a, Rodriguez et al., 2016) and the proportion of explained common variance (Sijtsma, 1998). Such proposals were supported in the current study in relation to person location estimation (but were not supported in relation to item discrimination estimation). Furthermore, in the current study, bias in the person location estimates, at the more extreme levels of the latent trait, also systematically increased with increase in the strength of LD (correlation of 0.59 and

0.69 between bias and strength of LD at the low and high level of latent trait, respectively). Such results are consistent with some of the previous findings from the LD line of research, indicating importance of LD in distortions in person location estimation (Zenisky et al., 2002).

In terms of the size of bias, in most of the conditions bias was less than 0.5 points (on the standard normal scale); that is, person location was overestimated in the case of the low trait, and underestimated for the high level of trait for less than 0.5 SD points. In one condition, bias could be safely characterized as excessive – it exceeded 0.8 SD points at the low and high end of the traits. In this condition, the dominant factor was 'weak' (an average simulated item discrimination of 0.5, corresponding to accounting for about 10% of overall standardized variance in a measure). Such a condition would be unlikely in real measures in practical use, as the factor analytical solution with the dominant factor of such strength would be deemed invalid. The person location was misestimated for less than 0.25 SD, when the dominant dimension was very strong (average item discrimination of 1.9) and the LD was weak or negligible (average item discrimination on the secondary dimension less than 0.5). The importance of different amounts of bias, in the context of real-life measures, should be interpreted in the specific practical contexts.

In the current investigation, a short measure (15 items) was of interest. In the previous research, it was found that the effects of the violations of unidimensionality on ability estimation were greater in shorter measures compared to longer measures (Zhang, 2008). The correlation between the ability estimate and the true ability decreased and the error increased with decrease in the length of the measure (e.g., from 30 to 15 items), when violations of unidimensionality were present. Therefore, with longer measures, accuracy in the IRT estimation of person location parameter is expected to increase. Further evidence about the effect of the length of a measure in

different conditions of violations of unidimensionality/LI is needed. The effect of the method of person location estimation also need further examination.

5.3 Eigenvalues-Based Indexes and Distortions in IRT Model Estimates

In relation to the second goal of the study – an examination of the utility of the four eigenvalues-based indexes in predicting distortions in unidimensional model estimates, the results suggested limited value of use of eigenvalues for such a purpose. None of the indexes were strongly predictive of bias in item parameters estimation – the strongest relation (a correlation coefficient of 0.56, recorded between Index 1 and bias in item discrimination), was not sufficient for the investigated purpose. The strong relations, however, were found between three of the indexes, which primarily reflected the strength of the dominant dimension (i.e., Index 2, Index 3, and Index 4) and the accuracy in estimation of the person location parameter. Such results are in accordance with the findings from the first part of this study, which suggested that the size of the dominant dimension was the most important factor for the accuracy in person location estimation (at the more extreme levels of the latent trait). In estimation of person location at the more extreme levels of the latent trait, the three indexes demonstrated possible value: correlations between the indexes and bias ranged from 0.82 to 0.89 based on Pearson correlation matrix, and from 0.75 to 0.87 based on tetrachoric correlation matrix.

In terms of the accuracy in item parameters estimation, for which, based on the findings from the first part of the study, the strength of LD was of greater importance, none of the indexes were useful. Poor utility of eigenvalues in indexing the strength of the factors beyond the first one, in common-factor model and with complex items, was indicated in the earlier research (Zopluoglu & Davenport, 2017). Increase in the size of the first eigenvalue on account of the size of the second eigenvalue was discussed and demonstrated in Zopluoglu & Davenport's study. In

relation to this, the ratio of the first to second eigenvalue did not demonstrate strong relations (i.e., strong enough for the purpose investigated in this study) with bias in estimation of item parameters or person location parameter, in the type of violations simulated in this study. The noted problems with indexing variance associated with the factors beyond the first one were reflected in significant variability/instability in estimation of the second eigenvalues (and in indexes that utilized the second eigenvalue along the first eigenvalue). The problems, likely related to the use of correlation matrix with communalities on the diagonal, were pronounced in tetrachoric correlation matrix, for which different types of estimation in EFA have been recommended (e.g., the robust weighted least squares estimator; Muthén, du Toit, & Spisic, 1997).

About the use of tetrachoric and Pearson correlations matrices, some differences were noted in regard to the magnitude of the relations between the indexes and bias; however, in cases when the strong relations with bias were recorded, the same conclusions were made based on the use of both types of indexes (i.e., both types of correlation matrices). The differences recorded in the magnitude of the calculated eigenvalues/eigenvalues-based indexes (i.e., they were consistently larger in Pearson correlation matrix), on the other hand, would impact the specific practical recommendations if the indexes are used – different specific recommendations would be given based on the use of the two matrices. The first eigenvalue was more accurate (i.e., closer to the true generated parameters) in the tetrachoric correlation matrix, consistent with the findings that Pearson's correlations represent underestimation when the data are categorical (West, Finch & Curran, 1995).

The focus of the current study was on weaker forms of assumptions violations – eigenvalues and eigenvalues-based indexes may perform differently in assessing different types

of violations, such as those when "substantive" secondary factors underlie item responses, which were not investigated in this study. It should be noted that in the data structures simulated in the current study, the "eigenvalue >1" rule suggested that one factor should be retained in all conditions[27] (consistent with how the conditions were conceptualized, as essentially unidimensional conditions with some LD present). However, some conditions of violations, in which bias in item discrimination was substantial, were not identified as problematic based on the use of the "eigenvalue >1" rule. The use of eigenvalues for the purpose of deciding how many dimensions to retain in the model is related to but distinct from the use that was of main interest in the current study.

5.4 Limitation and Future Directions

In the conditions simulated in the current study, intended to depict different levels of violation of the two assumptions, there was a smaller variability in simulated true item parameters compared to variability in the real-life individual measures. Less diverse item parameters were simulated in the current study in order to depict and study violations of certain type, i.e., to avoid the types of violations not of interest in the current study, such as those that may be reflective of substantive multidimensionality in a measure. Applicability of the obtained results to the measures with greater variability in the item parameters should be assessed.

In the model used in the study, various sources of LD were represented by a single nuisance dimension in order to provide a simpler context for studying the violations. In real-life measures, several specific sources of LD may be relevant and several secondary dimensions should be posited. How much such misspecification is relevant should be a subject of further

[27] None of the second eigenvalues in the simulated conditions were greater than 1.

examination. In a research context, several weaker secondary dimensions were associated to smaller distortions in unidimensional model estimates compared to presence of a single stronger dimension (Zhang, 2008); therefore, the distortions/bias found in the conditions designed in this study could be seen as the 'maximum' amount in such a context (i.e., distortions/bias are likely smaller if several weaker secondary dimensions are of actual relevance). In the within-item multidimensionality context, however, estimating models with multiple dimensions is challenging methodologically; moreover, conceptually, it is difficult to determine precisely what the items of our measures unintendedly tap into and to distinguish multiple sources of common variance that is not due to the construct we intended to measure.

In terms of creating different conditions of violation of assumptions, only certain points of the strength of the underlying dimension, from the continuum of all possible values, were selected. The selected values were based on the criteria utilized in the IRT literature (Baker, 2001), and the ten conditions of violations were created as an appropriate starting point in this investigation. In accordance with the conceptualization of the conditions simulated in the current study, random-effects design would be appropriate. In this study, however, the same set of item parameters were used across replication while only person location parameters were treated as random (simulated in each replication). This was done in order to keep complex/computationally demanding simulation design more manageable. Furthermore, due to the novelty of the used research design and exploratory nature of the study, the focus of the study was on the variations 'across' the simulated conditions, with less importance given to the precision 'within' the conditions. The current study is intended as a first step in this research direction – a random item parameter generation and a greater attention to 'within' conditions estimation will be

implemented in the following steps. More specific recommendations, regarding the amounts of the violations at the different levels of distortions, would be provided based on such a design.

Finally, the investigation conducted in current study was done with normally distributed latent traits. Further investigation of the research problem of interest should be conducted in the context of various types and degrees of violations of normality, relevant for different types of psychological measures – due to the scope of the current project, such conditions were not addressed at this time. A relatively short measure was of interest in the current investigation - further investigation about the effects of the length of measure, at different levels of violations of the two assumptions, on the estimation of different model parameters is needed.

5.5 Summary and Conclusions

The specific study design and the model utilized in the study proved useful in providing more details about the relations between the violations of the unidimensionality and LI assumptions and the accuracy in unidimensional IRT models. In the context of the type of violations simulated in the study (weaker forms of violations, shorter measures, greater item complexity) and the unidimensional IRT model used (2PL model), a general conclusion about the robustness of unidimensional IRT models was supported: the smaller violations of the assumptions, the less serious consequences for the accuracy of the unidimensional IRT model. Greater violations of the assumptions led to more serious consequences.

The evidence was provided about a strong relation between increase in the strength of LD and the size of distortions in the item parameters estimates. One of the consequences of LD in the measures intended to measure one construct, found in the study, is an overestimation of item discrimination parameter (i.e., factor loadings). Overestimation of item discrimination leads to overestimation in the precision of the measure and inflation in the researchers' confidence in the

conclusions based on the use of the measure. Furthermore, bias in item discrimination and location affects consequent analyses and structural-model coefficients based on the use of estimated parameters, as well as other uses of the parameters, specific for the IRT context.

For researchers interested in estimation of person location, the results of the study indicated that the estimates of person location were free of bias in the middle range of latent trait (i.e., for the majority of the respondents), with precision in the estimates decreasing with decrease in the strength of the dominant dimension and with increase in the strength of LD. At the more extreme ends of the latent trait, bias was found (i.e., overestimation of person location at the lower end and underestimation at the higher end of latent trait). The consequences of bias in person location, at the more extreme ends of latent trait, could be important. For example, when the cut points are dividing the test score range into two or more ordered categories, reflecting screening or diagnostic decisions at the more extreme ends of the distribution, overestimation at the low end and underestimation at the high end of the distribution are relevant. That is, the underestimation/overestimation of person location may influence the decisions based on the use of such cut points. The importance of the underestimation/overestimation is related to the types of the decisions and the seriousness of the possible consequences (i.e., underestimation at the high end of the latent trait would be of greater relevance in the context of a forensic psychology measure compared to an educational psychology measure).

Regarding the magnitude of bias in person location estimates reported in the current study, as the size of bias is also dependent on the length of the measure, the findings based on the particular length of measure used in the study (15 items) should not be generalized to either longer nor shorter measures. More information regarding the effects of the length of the measure in the context of different amounts of violations of the assumptions (e.g., how short measure

could be for satisfactory person location estimation) and in the context of different methods of estimation is needed. It is also important to keep in mind that the IRT model assessed in the current study was a two-parameter IRT model. As the parameters estimated from different IRT models are not directly comparable, the findings should not be generalized to one-parameter or three-parameter IRT models.

In relation to the investigation of the ways to provide the recommendations about the amount of distortions at different levels of violations, limited utility of eigenvalues-based indexes was found. None of the indexes demonstrated value in predicting bias in item parameters (for which the strength of LD was of importance). Other indexes, such as the IRT unidimensionality indexes and those based on different methods of estimation in FA (Nandakumar & Stout, 1993; Rodriguez et al., 2016; Zhang & Stout, 1999), should be further investigated for this purpose. The eigenvalues-based indexes primarily based on the size of the first eigenvalue (reflecting the strength of the dominant factor in a measure) demonstrated a possible value in relation to estimation of person location. The higher the indexes (i.e., Index 2, 3 and 4), the smaller distortions in person parameter estimation can be expected at the more extreme ends of the latent trait. More detailed analysis of such indexes should be pursued if distortions in person location estimation are of interest. Whether there is a single index that can be used for predicting bias in both item parameters and person location parameter needs further investigation.

Based on the results of the current study, one general rule that is relevant for both item parameters and person location parameter estimation could be stated as follows: When the

dominant factor accounted for at least 50% of overall standardized variance[28] and when LD was negligible or weak[29], the maximum bias in item discrimination was less than 0.1, the maximum bias in item location was less than 0.2 SD, and the maximum bias in the person location estimates was about 0.25 SD (at the more extreme levels of the latent trait). Such magnitude of distortions in the unidimensional estimates could be characterized as not substantial for many practical applications. With increase in the strength of LD, the accuracy in the estimation of both item discrimination and person location decreased. With decrease in the strength of the dominant dimension, the accuracy in the estimation of person location was primarily affected. Based on the data provided in the study, the guidelines for different levels of violations and the size of the distortions in the model estimates (estimates of item parameters and person location parameter) could be formed.

Overall, the findings of the current study highlighted the importance of following the recommended psychometric practices in the construction and use of measures and items. Well-designed measures intended to measure a single construct are characterized by a strong general factor *and* negligible/weak LD. Both elements are of importance, as demonstrated in the current study. Constructing and selecting the items with strong relations with the latent constructs is important in psychometric practice and practical research. In real-life measures, items typically reflect "something else" in addition to the construct the measure was intended to measure (i.e., various sources of construct-irrelevant variance); however, adequate steps should be taken to

[28] Indicated by Index 3 based on tetrachoric correlation matrix (calculated from the data, Table 14) and by the true parameters used in the simulation (average item discrimination of 1.9 simulated in these conditions).
[29] Average item discrimination of 0.5 and less was simulated on the secondary dimension for negligible or weak LD conditions (true parameters used in simulation).

minimize such sources, in the process of measure construction and with each use of the measure. With increase in "multidimensionality" in a measure intended to measure a single construct, the use of particular multidimensional models, as a way of accounting for LD in a measure, may be needed (Ip & Chen, 2012; Kaharman, 2013; Reise et al., 2013; Wainer et al., 2007).

References

Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*, 113–127. doi:10.1177/014662168901300201

Alessandri, G., Vecchione, M., Eisenberg, N., & Łaguna, M. (2015). On the factor structure of the Rosenberg (1965) General Self-Esteem Scale. *Psychological Assessment, 27*, 621–635. doi:10.1037/pas0000073

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement, 9*, 37–48. doi: 10.1177/014662168500900104

Baker, F. (2001). *Basics of item response theory.* Retrieved from http://echo.edres.org:8080/irt/baker

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison–Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29–51. doi: 10.1007/BF02291411

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal, 22*, 504–516. doi:10.1080/10705511.2014.938596

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–424). Reading, MA: Addison–Wesley.

Bulut, O., & Sünbül, Ö. (2017). Monte Carlo simulation studies in item response theory with the R programming language. *Journal of Measurement and Evaluation in Education and Psychology, 8,* 266–287. doi: 10.21031/epod.305821

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245–276. doi:10.1207/s15327906mbr0102_10

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48,* 1–29. doi:10.18637/jss.v048.i06

Chen, W. H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics, 22*, 265–289. doi: 10.2307/1165285

Cho, S. J., Li, F., & Bandalos, D. (2009). Accuracy of the parallel analysis procedure with polychoric correlations. *Educational and Psychological Measurement, 69*, 748–759. doi:10.1177/0013164409332229

Crişan, D. R., Tendeiro, J. N., & Meijer, R. R. (2017). Investigating the practical consequences of model misfit in unidimensional IRT models. *Applied Psychological Measurement, 41,* 439–455. doi: 10.1177/0146621617695522

de Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36,* 104–121. doi.org/10.1177/0146621612437403

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. *Applied Psychological Measurement, 7,* 189–199. doi/10.1177/014662168300700207

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

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*, 36–49. doi.org/10.1111/emip.12111

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466–491. doi: 10.1037/1082-989X.9.4.466

Folk, V. G., & Green, B. F. (1989). Adaptive estimation when the unidimensionality assumption of IRT is violated. *Applied Psychological Measurement*, *13*, 373–389. doi:10.1177/014662168901300404

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: Erlbaum.

Hambleton, R. K., & Murray, L. N. (1983). Some goodness of fit investigations for item response models. In R. K. Hambleton (Ed.), *Applications of item response theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer.

Harrison, D. A. (1986). Robustness of IRT parameter estimation to violations of the unidimensionality assumption. *Journal of Educational Statistics, 11*, 91–115. doi:10.2307/1164972

Harwell, M., Stone, C. A., Hsu, T. C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*, 101–125. doi:10.1177/014662169602000201

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185. doi:10.1007/BF02289447

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Ip, E. H. (2010). Empirically indistinguishable multidimensional IRT and locally dependent unidimensional item response models. *British Journal of Mathematical and Statistical Psychology, 63,* 395–416. doi: 10.1348/000711009X466835

Ip, E. H. S., & Chen, S. H. (2012). Projective item response model for test-independent measurement. *Applied Psychological Measurement, 36,* 581–601. doi: 10.1177/0146621612452778

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36,* 347–387. doi: 10.1207/S15327906347-387.

Kahraman, N. (2013). Unidimensional interpretations for multidimensional test items. *Journal of Educational Measurement, 50,* 227–246. doi:10.1111/jedm.12012

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurements, 20,*141–151.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling, 15,* 136–153. doi: 10.1080/10705510701758406

Kirisci, L., Hsu, T. C., & Yu, L. (2001). Robustness of item parameter estimation programs to assumptions of unidimensionality and normality. *Applied Psychological Measurement, 25*, 146–162. doi:10.1177/01466210122031975

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction* (pp. 362–412). New York, NY: Wiley.

Lind, J. C., & Zumbo, B. D. (1993). The continuity principle in psychological research: An introduction to robust statistics. *Canadian Psychology/Psychologie Canadienne, 34*, 407–414. doi: 10.1037/h0078861

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph No. 7. Richmond, VA: Psychometric Corporation. Retrieved from: http://www.psychometrika.org/journal/online/MN07.pdf

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. New York: Erlbaum Associates.

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

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment, 22*, 366–381. doi: 10.1037/a0019225

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34*, 100–117.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement, 24,* 99–114. doi:10.1177/01466210022031552

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50, 741-749.

Mislevy, R. J. (1986). Recent developments in the factor analysis of categorical variables. *Journal of Educational Statistics 11,* 3–31. doi: 10.2307/1164846

Mislevy, R.J., & Bock, R.D. (1983). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville IN: Scientific Software

Mulaik, S. A. (2009). *Foundations of factor analysis.* Chapman and Hall/CRC.

Muthén, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics, 22,* 48–65. doi: 10.1016/0304-4076(83)90093-3

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes* (Technical Report). Los Angeles, CA: University of California.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement*, *28,* 99–117. doi: 10.1111/j.1745-3984.1991.tb00347.x

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics*, *18,* 41–68. doi: 10.3102/10769986018001041

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika, 44,* 443–460. doi: 10.1007/BF02296207

R Development Core Team (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available from https://www.R-project.org/

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen, Denmark: Danish Institute for Educational Research.

Revelle, W. (2017). *psych: Procedures for psychological, psychometric, and personality research*. Evanston, IL: Northwestern University.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4,* 207–230. doi: 10.3102/10769986004003207

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013a). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment, 95,* 129–140. doi: 10.1080/00223891.2012.725437

Reise, S. P., & Rodriguez, A. (2016). Item response theory and the measurement of psychiatric constructs: Some empirical and conceptual issues and challenges. *Psychological Medicine, 46,* 2025–2039. doi: 10.1017/S0033291716000520

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013b). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73,* 5–26. doi: 10.1177/0013164412449831

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98,* 223–237. doi: 10.1080/00223891.2015.1089249

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25. doi: 10.18637/jss.v017.i05

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*, 282–292. doi: 10.1037/a0025697

Sahin, S. G., Walker, C. M., & Gelbal, S. (2015). The impact of model misspecification with multidimensional test data. In *Quantitative Psychology Research* (pp. 145–172). Springer, Cham.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107–120. doi:10.1007/s11336-008-9101-0

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237–247.

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the unidimensionality of psychological scales: Using multiple criteria from factor analysis. *Social Indicators Research, 102,* 443–461. doi: 2011-11661-006

Slocum-Gori, S. L., Zumbo, B. D., Michalos, A. C., & Diener, E. (2009). A note on the dimensionality of quality of life scales: An illustration with the Satisfaction with Life Scale (SWLS). *Social Indicators Research*, *92*, 489–496.doi: 10.1007/s11205-008-9303-y

Steinberg, L., & Thissen, D. (1996). Uses of item response theory and the testlet concept in the measurement of psychopathology. *Psychological Methods*, *1*, 81–97. doi: 10.1037/1082-989X.1.1.81

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589–617. doi: 10.1007/BF02294821

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408. doi: 10.1007/BF02294363

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of responses to test items. *Applied Psychological Measurement*, *27*, 159–203. doi: 10.1177/0146621603027003001

Traub, R. (1983). A prior consideration in choosing an item response model. In R. K. Hambleton (Ed.), *Applications of Item Response Theory*. Vancouver, BC: Educational Research Institute of British Columbia.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explicationthrough factor or component analysis: A review and evaluation ofalternative procedures for determining the number of factors or compo-nents. In E. Helmes (Ed.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41–71). New York, NY: Kluwer Academic/Plenum.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press. doi:10.1017/CBO9780511618765

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15*, 22–29. doi: 10.1111/j.1745-3992.1996.tb00803.x

Wang, W., Cheng, Y., & Wilson, M. (2005). Local item dependence for items across tests connected by common stimuli. *Educational and Psychological Measurement, 65*, 5–27. doi: 10.1177/0013164404268676

Way, W. D., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement, 12*, 239–252. doi: 10.1177/014662168801200303

Weng, L., & Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*, 697–716. doi:10.1177/0013164404273941

West, S. G, Finch, J. F., & Curran, P. J. (1995). Structural equation models with non-normal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56−75). Thousand Oaks, CA: Sage.

Wingersky, M.S., Barton, M.A., & Lord, F.M. (1982). LOGIST user's guide. Princeton NJ: Educational Testing Service

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods, 12*, 58–79. doi: 10.1037/1082-989X.12.1.58

Yang, Y., & Xia, Y. (2015). On the number of factors to retain in exploratory factor analysis for ordered categorical data. *Behavior Research Methods, 47*, 756–772. doi: 10.3758/s13428-014-0499-2

Yen, W. M. (1984). Effects of item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145. doi: 10.1177/014662168400800201

Yen, W. M. (1987). A comparison of the efficiency and accuracy of BILOG and LOGIST. *Psychometrika, 50,* 275–291. doi: 1988-00192-001

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30,* 187–213. doi: 10.1111/j.1745-3984.1993.tb00423.x

Zenisky, A. L., Hambleton, R. K., & Sired, S. G. (2002). Identification and evaluation of local item dependencies in the Medical College Admissions Test. *Journal of Educational Measurement, 39,* 291–309.doi: j.1745-3984.2002.tb01144.x

Zhang, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimensions. *The Journal of Experimental Education, 77,* 147–166. doi: 10.3200/JEXE.77.2.147-166

Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika, 64,* 213–249. doi: 10.1007/BF02294536

Zopluoglu, C., & Davenport Jr, E. C. (2017). A note on using eigenvalues in dimensionality assessment. *Practical Assessment, Research & Evaluation, 22.*

Appendix A

Item Discrimiantion and Item Location Parameters Simulated in the Study

Table 1A

Item Discrimination and Item Location Parameters for Condition 1 (Very Strong – Negligible)

Items	a_1	a_2	b_1	b_2
Item 1	1.66	0.25	0.44	0.15
Item 2	2.03	0.35	-0.95	-1.66
Item 3	1.76	0.12	0.08	0.46
Item 4	1.78	0.21	0.19	-1.86
Item 5	1.71	0.14	-0.31	0.81
Item 6	1.96	0.03	-0.18	-0.68
Item 7	1.88	0.23	0.82	-0.89
Item 8	1.78	0.36	0.61	-1.39
Item 9	2.15	0.23	-1.14	-1.49
Item 10	1.92	0.21	1.55	0.28
Item 11	1.7	0.02	-1.08	-0.87
Item 12	1.87	0.09	0.5	1.71
Item 13	2.3	0.19	-1.65	-0.6
Item 14	1.84	0.22	2.01	0.16
Item 15	1.78	0	0.06	-0.92
M (SD)	1.87 (0.18)	0.18 (0.11)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 2A

Item Discrimination and Item Location Parameters for Condition 2 (Very Strong – Weak)

Items	a_1	a_2	b_1	b_2
Item 1	1.66	0.39	0.44	0.15
Item 2	2.03	0.55	-0.95	-1.66
Item 3	1.76	0.68	0.08	0.46
Item 4	1.78	0.41	0.19	-1.86
Item 5	1.71	0.4	-0.31	0.81
Item 6	1.96	0.49	-0.18	-0.68
Item 7	1.88	0.34	0.82	-0.89
Item 8	1.78	0.73	0.61	-1.39
Item 9	2.15	0.42	-1.14	-1.49
Item 10	1.92	0.53	1.55	0.28
Item 11	1.7	0.6	-1.08	-0.87
Item 12	1.87	0.5	0.5	1.71
Item 13	2.3	0.61	-1.65	-0.6
Item 14	1.84	0.49	2.01	0.16
Item 15	1.78	0	0.06	-0.92
M (SD)	1.87 (0.18)	0.48 (0.17)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 3A

Item Discrimination and Item Location Parameters for Condition 3 (Very Strong – Moderate)

Items	a_1	a_2	b_1	b_2
Item 1	1.66	0.49	0.44	0.15
Item 2	2.03	1.07	-0.95	-1.66
Item 3	1.76	0.95	0.08	0.46
Item 4	1.78	0.99	0.19	-1.86
Item 5	1.71	1.19	-0.31	0.81
Item 6	1.96	0.8	-0.18	-0.68
Item 7	1.88	1.29	0.82	-0.89
Item 8	1.78	1.17	0.61	-1.39
Item 9	2.15	1.17	-1.14	-1.49
Item 10	1.92	1.33	1.55	0.28
Item 11	1.7	0.75	-1.08	-0.87
Item 12	1.87	0.6	0.5	1.71
Item 13	2.3	0.9	-1.65	-0.6
Item 14	1.84	1.07	2.01	0.16
Item 15	1.78	0	0.06	-0.92
M (SD)	1.87 (0.18)	0.92 (0.35)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 4A

Item Discrimination and Item Location Parameters for Condition 4 (Very Strong – Strong)

Items	a_1	a_2	b_1	b_2
Item 1	1.66	1.58	0.44	0.15
Item 2	2.03	1.28	-0.95	-1.66
Item 3	1.76	1.48	0.08	0.46
Item 4	1.78	1.47	0.19	-1.86
Item 5	1.71	1.54	-0.31	0.81
Item 6	1.96	1.33	-0.18	-0.68
Item 7	1.88	1.39	0.82	-0.89
Item 8	1.78	1.47	0.61	-1.39
Item 9	2.15	1.21	-1.14	-1.49
Item 10	1.92	1.36	1.55	0.28
Item 11	1.7	1.54	-1.08	-0.87
Item 12	1.87	1.40	0.5	1.71
Item 13	2.3	1.12	-1.65	-0.6
Item 14	1.84	1.42	2.01	0.16
Item 15	1.78	0	0.06	-0.92
M (SD)	1.87 (0.18)	1.30 (0.38)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 5A

Item Discrimination and Item Location Parameters for Condition 5 (Strong – Negligible)

Items	a_1	a_2	b_1	b_2
Item 1	1.58	0.25	0.44	0.15
Item 2	1.28	0.35	-0.95	-1.66
Item 3	1.48	0.12	0.08	0.46
Item 4	1.47	0.21	0.19	-1.86
Item 5	1.54	0.14	-0.31	0.81
Item 6	1.33	0.03	-0.18	-0.68
Item 7	1.39	0.23	0.82	-0.89
Item 8	1.47	0.36	0.61	-1.39
Item 9	1.21	0.23	-1.14	-1.49
Item 10	1.36	0.21	1.55	0.28
Item 11	1.54	0.02	-1.08	-0.87
Item 12	1.40	0.09	0.5	1.71
Item 13	1.12	0.19	-1.65	-0.6
Item 14	1.42	0.22	2.01	0.16
Item 15	1.47	0	0.06	-0.92
M (SD)	1.40 (0.13)	0.18 (0.11)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 5A

Item Discrimination and Item Location Parameters for Condition 6 (Strong – Weak)

Items	a_1	a_2	b_1	b_2
Item 1	1.58	0.39	0.95	1.33
Item 2	1.28	0.55	-1.15	1.26
Item 3	1.48	0.68	-2.16	1.36
Item 4	1.47	0.41	0.97	0.64
Item 5	1.54	0.40	-0.84	-0.68
Item 6	1.33	0.49	-1.23	-0.81
Item 7	1.39	0.34	-1.54	-1.04
Item 8	1.47	0.73	1.01	-0.1
Item 9	1.21	0.42	-0.85	2.15
Item 10	1.36	0.53	-2.90	-0.95
Item 11	1.54	0.60	1.48	0.01
Item 12	1.40	0.50	-0.52	-0.59
Item 13	1.12	0.61	0	0.01
Item 14	1.42	0.49	-0.27	-1.61
Item 15	1.47	0	1.90	0.33
M (SD)	1.40 (0.13)	0.48 (0.17)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 7A

Item Discrimination and Item Location Parameters for Condition 7 (Strong – Moderate)

Items	a_1	a_2	b_1	b_2
Item 1	1.58	0.49	0.44	0.15
Item 2	1.28	1.07	-0.95	-1.66
Item 3	1.48	0.95	0.08	0.46
Item 4	1.47	0.99	0.19	-1.86
Item 5	1.54	1.19	-0.31	0.81
Item 6	1.33	0.8	-0.18	-0.68
Item 7	1.39	1.29	0.82	-0.89
Item 8	1.47	1.17	0.61	-1.39
Item 9	1.21	1.17	-1.14	-1.49
Item 10	1.36	1.33	1.55	0.28
Item 11	1.54	0.75	-1.08	-0.87
Item 12	1.40	0.6	0.5	1.71
Item 13	1.12	0.9	-1.65	-0.6
Item 14	1.42	1.07	2.01	0.16
Item 15	1.47	0	0.06	-0.92
M (SD)	1.40 (0.13)	0.92 (0.35)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 8A

Item Discrimination and Item Location Parameters for Condition 8 (Moderate – Negligible)

Items	a_1	a_2	b_1	b_2
Item 1	0.49	0.25	0.44	0.15
Item 2	1.07	0.35	-0.95	-1.66
Item 3	0.95	0.12	0.08	0.46
Item 4	0.99	0.21	0.19	-1.86
Item 5	1.19	0.14	-0.31	0.81
Item 6	0.8	0.03	-0.18	-0.68
Item 7	1.29	0.23	0.82	-0.89
Item 8	1.17	0.36	0.61	-1.39
Item 9	1.17	0.23	-1.14	-1.49
Item 10	1.33	0.21	1.55	0.28
Item 11	0.75	0.02	-1.08	-0.87
Item 12	0.6	0.09	0.5	1.71
Item 13	0.9	0.19	-1.65	-0.6
Item 14	1.07	0.22	2.01	0.16
Item 15	1.17	0	0.06	-0.92
M (SD)	1.00 (0.25)	0.18 (0.11)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Table 9A

Item Discrimination and Item Location Parameters for Condition 9 (Moderate – Weak)

Items	a_1	a_2	b_1	b_2
Item 1	0.49	0.39	0.44	0.15
Item 2	1.07	0.55	-0.95	-1.66
Item 3	0.95	0.68	0.08	0.46
Item 4	0.99	0.41	0.19	-1.86
Item 5	1.19	0.4	-0.31	0.81
Item 6	0.8	0.49	-0.18	-0.68
Item 7	1.29	0.34	0.82	-0.89
Item 8	1.17	0.73	0.61	-1.39
Item 9	1.17	0.42	-1.14	-1.49
Item 10	1.33	0.53	1.55	0.28
Item 11	0.75	0.6	-1.08	-0.87
Item 12	0.6	0.5	0.5	1.71
Item 13	0.9	0.61	-1.65	-0.6
Item 14	1.07	0.49	2.01	0.16
Item 15	1.17	0	0.06	-0.92
M (SD)	1.00 (0.25)	0.48 (0.17)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

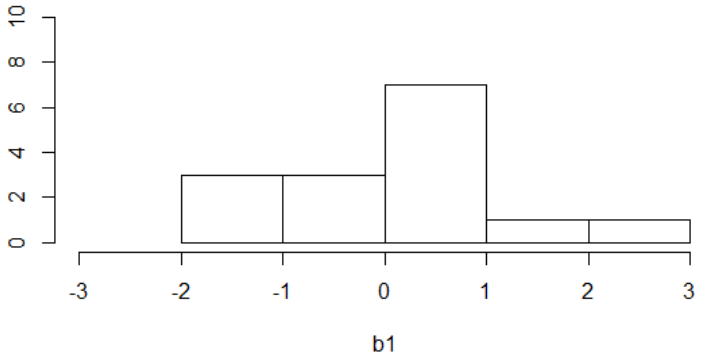
Table 10A

Item Discrimination and Item Location Parameters for Condition 10 (Weak – Negligible)

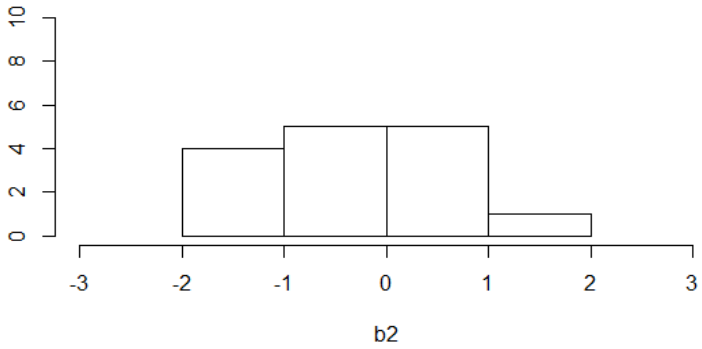
Items	a_1	a_2	b_1	b_2
Item 1	0.39	0.25	0.44	0.15
Item 2	0.55	0.35	-0.95	-1.66
Item 3	0.68	0.12	0.08	0.46
Item 4	0.41	0.21	0.19	-1.86
Item 5	0.4	0.14	-0.31	0.81
Item 6	0.49	0.03	-0.18	-0.68
Item 7	0.34	0.23	0.82	-0.89
Item 8	0.73	0.36	0.61	-1.39
Item 9	0.42	0.23	-1.14	-1.49
Item 10	0.53	0.21	1.55	0.28
Item 11	0.6	0.02	-1.08	-0.87
Item 12	0.5	0.09	0.5	1.71
Item 13	0.61	0.19	-1.65	-0.6
Item 14	0.49	0.22	2.01	0.16
Item 15	0.58	0	0.06	-0.92
M (SD)	0.51 (0.11)	0.18 (0.11)	0.06 (1.01)	-0.45 (1.02)

a_1 - item discrimination on the dominant dimension; a_2 – item discrimination on the secondary dimension; b_1 - item difficulty on the dominant dimension; b_2 – item difficulty on the secondary dimension.

Histogram for simulated b_1 parameters:



Histogram for simulated b_2 parameters:

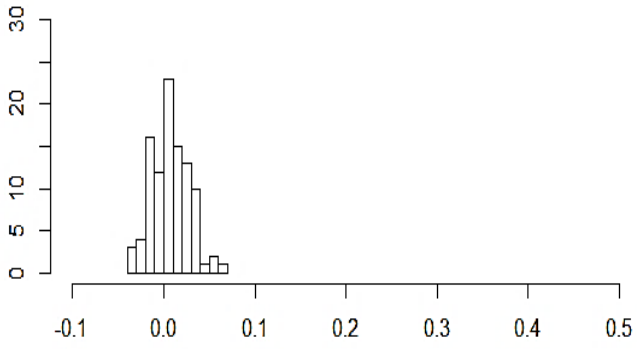


Appendix B

The Individual Graphs from the Figures from the Results Section.

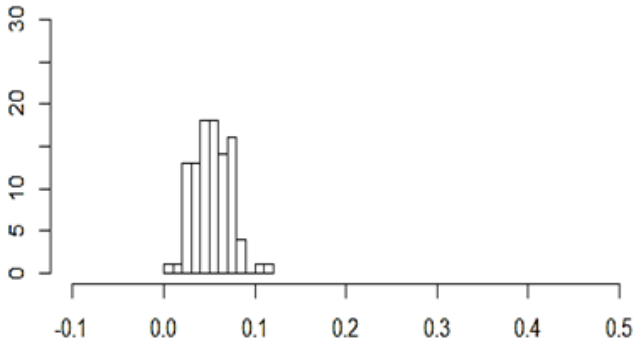
The Graphs from Figure 3:

Condition 1 (very strong – negligible)



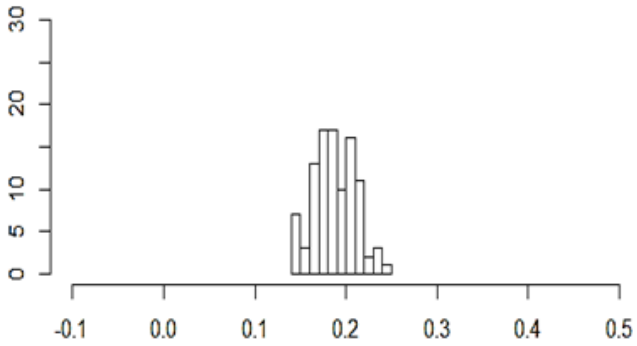
Bias in *a*

Condition 2 (very strong – weak)



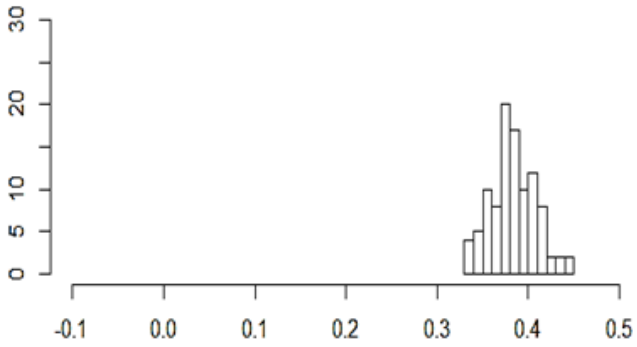
Bias in *a*

Condition 3 (very strong – moderate)



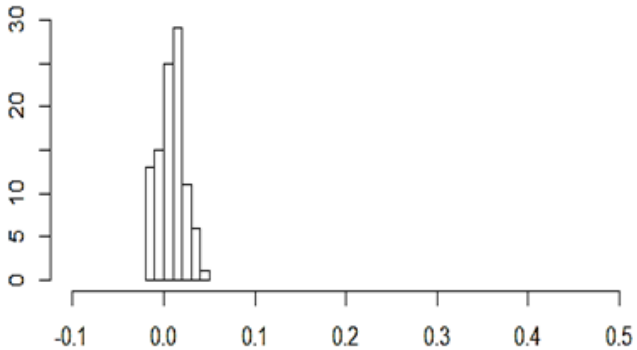
Bias in *a*

Condition 4 (very strong – strong)



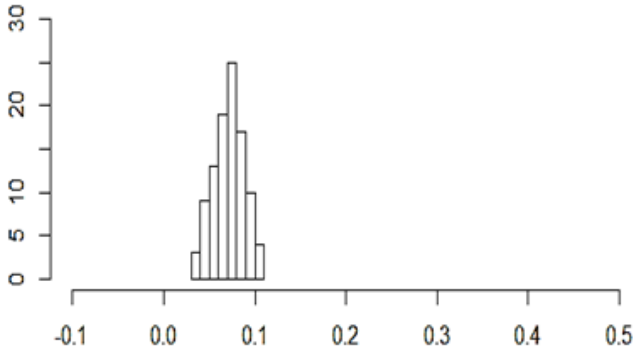
Bias in *a*

Condition 5 (strong – negligible)



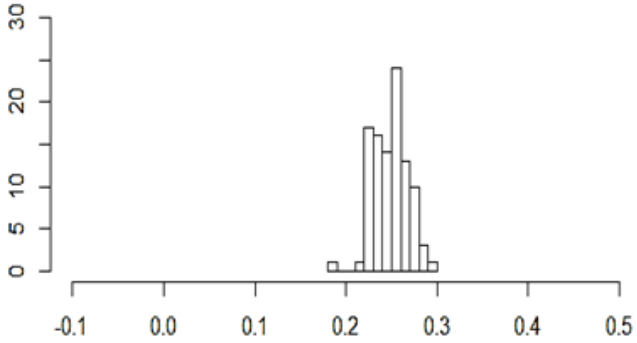
Bias in *a*

Condition 6 (strong – weak)



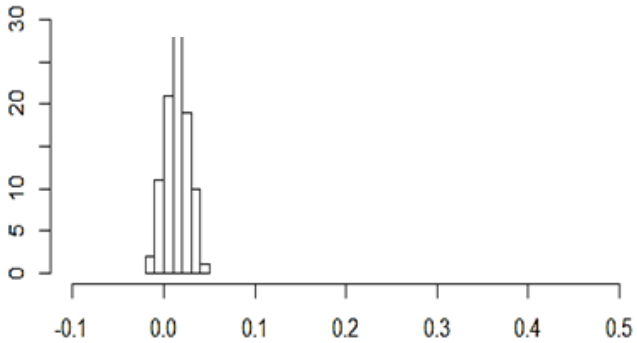
Bias in *a*

Condition 7 (strong – moderate)



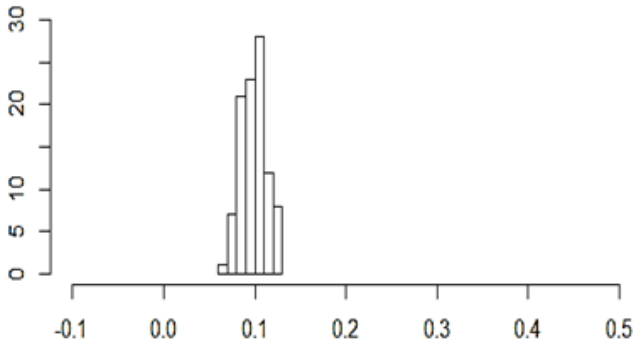
Bias in *a*

Condition 8 (moderate – negligible)



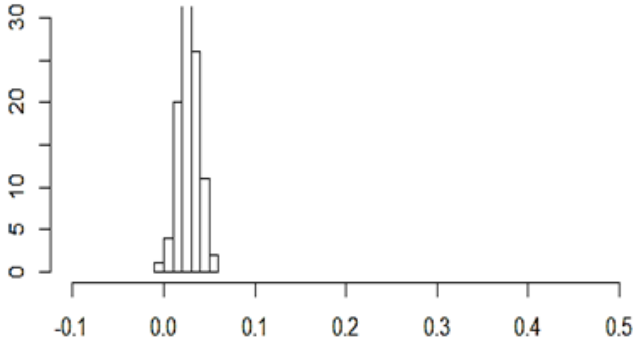
Bias in *a*

Condition 9 (moderate – weak)



Bias in *a*

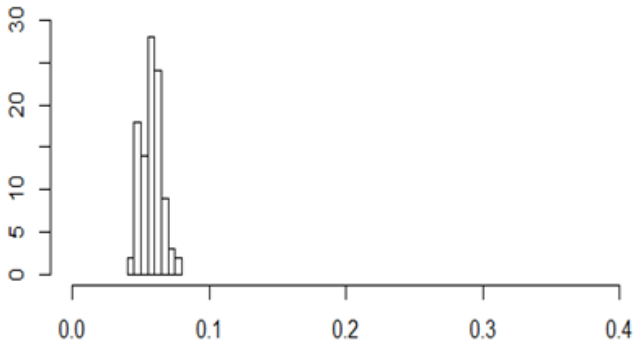
Condition 10 (weak – negligible)



Bias in *a*

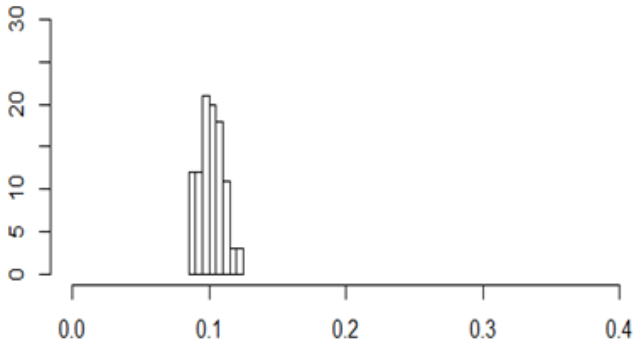
The Graphs from Figure 3, from the Results Section:

Condition 1 (very strong – negligible)



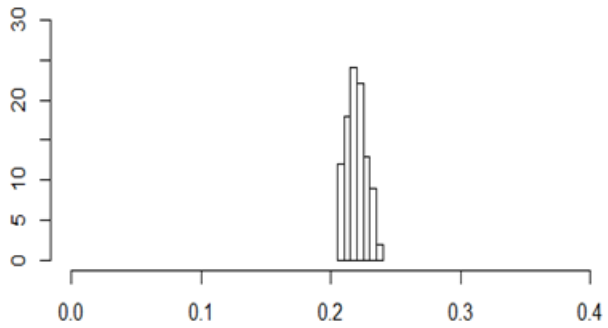
Bias in *b*

Condition 2 (very strong – weak)



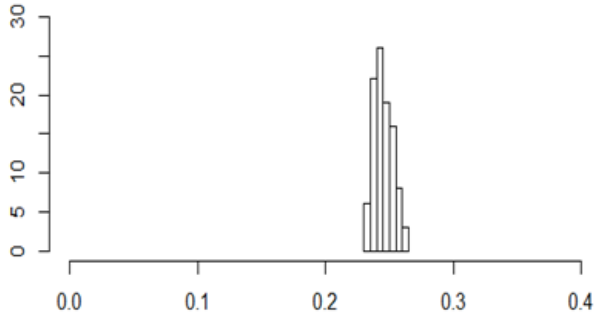
Bias in *b*

Condition 3 (very strong – moderate)



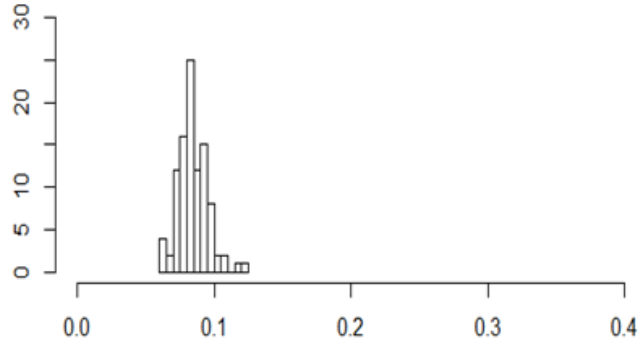
Bias in *b*

Condition 4 (very strong – strong)



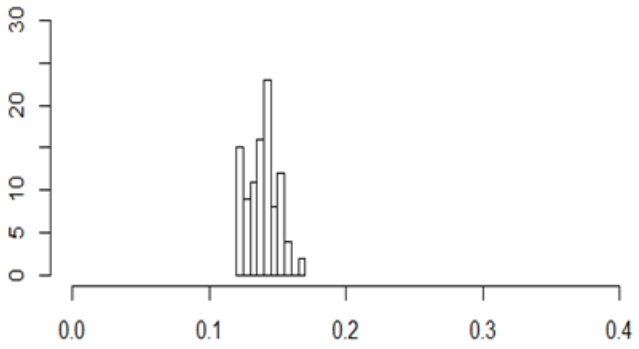
Bias in *b*

Condition 5 (strong – negligible)



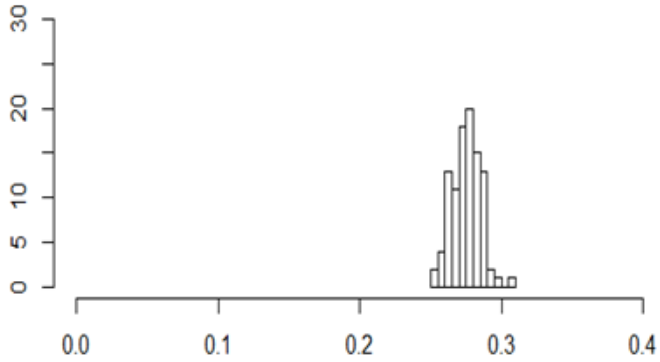
Bias in *b*

Condition 6 (strong – weak)



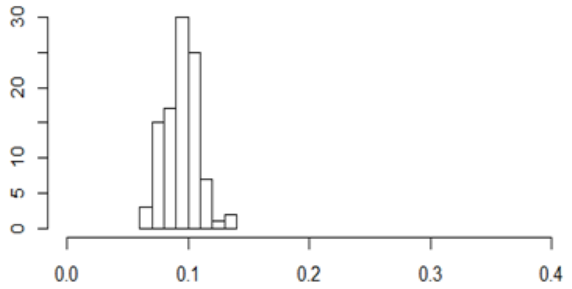
Bias in *b*

Condition 7 (strong – moderate)



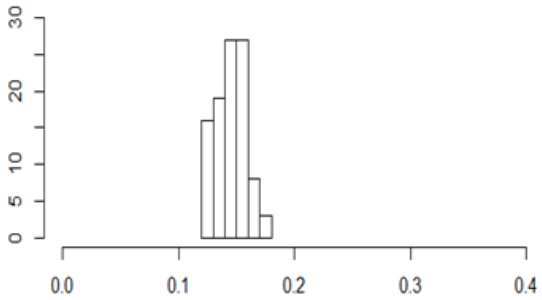
Bias in *b*

Condition 8 (moderate – negligible)



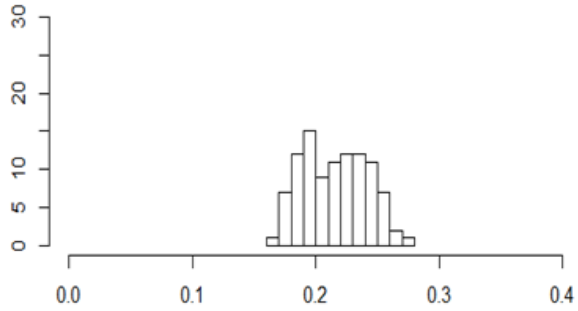
Bias in *b*

Condition 9 (moderate – weak)



Bias in *b*

Condition 10 (weak – negligible)



Bias in *b*