



ATHENS UNIVERSITY OF ECONOMICS AND  
BUSINESS

DEPARTMENT OF STATISTICS

---

**Factor analysis for daily diary PRO data:  
data handling decisions and their  
implications**

---

*Author:*  
Dumi Gerasimos

*Supervisor:*  
Karlis Dimitrios

**M.Sc. Thesis**

*Submitted to the Department of Statistics  
of the Athens University of Economics and Business  
in partial fulfilment of the requirements for  
the degree of Master of Science in Statistics*

Athens, Greece  
December 2023





# ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

## ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

---

Παραγοντική ανάλυση δεδομένων που  
συλλέγονται καθημερινά και  
αναφέρονται από τους ασθενείς: Στρατηγικές  
χειρισμού δεδομένων και επιπτώσεις

---

Συγγραφέας:  
Ντούμι Γεράσιμος

Επιβλέπων:  
Καρλής Δημήτριος

Μεταπτυχιακή Διατριβή  
Που υποβλήθηκε στο Τμήμα Στατιστικής  
του Οικονομικού Πανεπιστημίου Αθηνών  
ως μέρος των απαιτήσεων για την απόκτηση  
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα, Ελλάδα  
Δεκεμβριος 2023





# Acknowledgements

I would like to take this opportunity to express my sincere gratitude to my supervisor, Dimitris Karlis for his useful guidance and insightful discussions we had throughout the process of completing this thesis.

I am immensely grateful to Patient Centered Solutions (PCS) team at IQVIA for the internship opportunity they provided me and for assigning me an intriguing research topic for my thesis. I am also deeply grateful to the contributors to this work during my internship at IQVIA. Christina Daskalopoulou, Dara O'Neill and Pip Griffiths have graciously shared their expertise, providing valuable insights and feedback for this work.

Lastly, I would like to express my appreciation to my family and friends for their support and encouragement during this enormous journey.



## Abstract

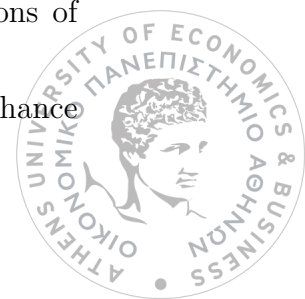
Daily diaries are an important modality for patient-reported outcome assessment. They typically comprise multiple questions, so understanding their underlying structure is key to appropriate analysis and interpretation. However, considering the high volume of repeated measurements, structural evaluation of such measures can pose challenges. Potential strategies include: (i) selecting a single day, (ii) averaging item-level observations over time, or (iii) using all data while accounting for its multilevel structure.

Using simulated diary datasets generated via a graded response model approach comprising correlated scores that emulate a 1 week period of data collection, the current study assessed the above strategies employing exploratory and confirmatory factor modelling. The assessment was primarily focused on the evaluation of the impact of each approach on different estimates including inter-item correlations, factor loading patterns, model fit (i.e., root mean square error of approximation, comparative fit index, Tucker Lewis index and standardized root mean residual) and the estimated the number of factors (i.e., Kaiser criterion, empirical Kaiser criterion and parallel analysis).

Both single day and item-average approaches resulted in biased factor loadings. The former displayed lower overall absolute average bias and overall mean square error, but greater frequency of incorrect factor count identification compared with the latter when using Kaiser criterion. The difference in the magnitude of bias and the mean square error between the single selected day and weekly item average approaches was higher in scenario 1 where the true parameter loadings were moderate (i.e., range: 0.62-0.70) compared to when they were high (i.e., range: 0.69-0.86). Increased inter-item correlations, relative to the simulated true values, were apparent in the item-average method. The root mean square error of approximation and Tucker Lewis index produced the most conservative results compared to the other goodness of fit measures. The standardized root mean residual almost invariably produced a good fit across all approaches under evaluation, even in the case of the item-average approach, where the estimated correlation matrix was divergent from the true correlation matrix under the hypothesized model.

The presence of non-trivial between- and within-individual variance highlighted the utility of a multilevel approach in examining the measurement properties of the diary instrument, such as dimensional and cross-level invariance. However, results also highlighted that convergence and Heywood cases can be more common with the multilevel approach, amongst low sample sizes typically encountered in clinical applications of such diary measures.

The findings established in this study suggest that a multilevel approach can enhance



the validity and utility of insight when evaluating the structural properties of daily diary data. However, there are still limitations under small sample size conditions. This multi-faceted investigation offers guidance on the impact of data handling decisions in diary assessment.



## Περίληψη

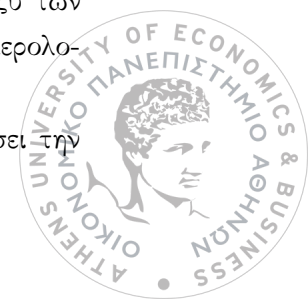
Τα καθημερινά ημερολογιακά δεδομένα είναι μια σημαντική μέθοδος συλλογής δεδομένων για την αξιολόγηση της έκβασης του ασθενούς που αναφέρεται από τον ίδιο τον ασθενή. Συνήθως περιλαμβάνουν πολλαπλές ερωτήσεις, επομένως η κατανόηση της υποκείμενης δομής τους είναι το κλειδί για την κατάλληλη ανάλυση και ερμηνεία. Ωστόσο, λαμβάνοντας υπόψη τον μεγάλο όγκο επαναλαμβανόμενων μετρήσεων, η δομική αξιολόγηση τέτοιων δεδομένων μπορεί να δημιουργήσει προκλήσεις. Οι πιθανές στρατηγικές περιλαμβάνουν: (i) την επιλογή μιας μεμονωμένης ημέρας, (ii) τη λήψη του μέσου όρου των παρατηρήσεων στον χρόνο για το κάθε άτομο ή (iii) τη χρήση όλων των δεδομένων, λαμβάνοντας υπόψη την πολυεπίπεδη δομή τους.

Χρησιμοποιώντας προσομοιωμένα ημερολογιακά δεδομένα μέσω ενός μοντέλου διαβαθμισμένης απόκρισης με συσχετισμένες λανθάνουσες μετρήσεις σε περίοδο 1 εβδομάδας, η παρούσα μελέτη αξιολόγησε τις παραπάνω στρατηγικές χρησιμοποιώντας διερευνητική και επιβεβαιωτική παραγοντική ανάλυση, αξιολογώντας το αντίκτυπο κάθε προσέγγισης σε διάφορες εκτιμήσεις, συμπεριλαμβανομένων των συσχετίσεων μεταξύ μετρήσεων, τα μοτίβα με βάση τις επιβαρύνσεις των παραγόντων, την προσαρμογή μοντέλου (δηλαδή, μέση τετραγωνική ρίζα σφάλματος προσέγγισης, συγκριτικός δείκτης προσαρμογής, Τούκερ Λούις δείκτης and τυποποιημένη ρίζα των αναμενόμενων καταλοίπων) και τον προσδιορισμό του αριθμού παραγόντων (δηλαδή, κριτήριο Kaiser, εμπειρικό κριτήριο Kaiser και παράλληλη ανάλυση).

Τόσο οι προσεγγίσεις μίας ημέρας όσο και οι προσεγγίσεις του μέσου όρου στοιχείων οδήγησαν σε μεροληπτικές επιβαρύνσεις των παραγόντων. Το πρώτο εμφάνισε χαμηλότερη συνολική απόλυτη μέση μεροληψία και συνολικό μέσο τετραγωνικό σφάλμα, αλλά μεγαλύτερη συχνότητα αναγνώρισης εσφαλμένου αριθμού παραγόντων σε σύγκριση με το δεύτερο όταν χρησιμοποιήθηκε το κριτήριο Kaiser. Το μέγεθος της μεροληψίας και το μέσο τετραγωνικό σφάλμα ήταν υψηλότερο για την προσέγγιση του μέσου όρου εβδομαδιαίων στοιχείων όταν η πραγματική επιβάρυνση ήταν μέτριου μεγέθους (δηλαδή, εύρος: 0.62-0.70) σε σύγκριση με όταν ήταν υψηλού μεγέθους (δηλαδή, εύρος: 0.69-0.86). Αυξημένες συσχετίσεις μεταξύ στοιχείων, σε σχέση με τις προσομοιωμένες πραγματικές τιμές, ήταν εμφανείς στη μέθοδο του μέσου όρου στοιχείων. Η μέση τετραγωνική ρίζα σφάλματος προσέγγισης και ο δείκτης Τούκερ Λούις παρήγαγαν τα πιο συντηρητικά αποτελέσματα σε σύγκριση με τα άλλα μέτρα καλής προσαρμογής. Η τυποποιημένη ρίζα των αναμενόμενων καταλοίπων σχεδόν πάντα παρήγαγε καλή προσαρμογή σε όλες τις υπό αξιολόγηση προσεγγίσεις, ακόμη και στην περίπτωση της προσέγγισης του μέσου όρου των στοιχείων όπου ο εκτιμώμενος πίνακας συσχέτισης ήταν αποκλίνων από τον πραγματικό πίνακα συσχέτισης κάτω από μοντέλο προσομοίωσης.

Η παρουσία σημαντικού ποσοστού διακύμανσης μεταξύ των ατόμων και εντός του ατόμου υπογράμμισε τη χρησιμότητα μιας πολυεπίπεδης προσέγγισης στην εξέταση των ιδιοτήτων μέτρησης του ημερολογιακού οργάνου. Ωστόσο, τα αποτελέσματα τόνισαν επίσης ότι η σύγκλιση και οι περιπτώσεις Χείγουντ μπορεί να είναι πιο κοινές με την πολυεπίπεδη προσέγγιση, μεταξύ των μεγεθών δειγμάτων που συνήθως χρησιμοποιούνται σε κλινικές εφαρμογές τέτοιων ημερολογιακών μετρήσεων.

Τα ευρήματά αυτά υποδηλώνουν ότι μια πολυεπίπεδη προσέγγιση μπορεί να ενισχύσει την



εγκυρότητα και τη χρησιμότητα των πληροφοριών που αφορούν την αξιολόγηση των δομικών ιδιοτήτων των ημερολογιακών δεδομένων. Ωστόσο, εξακολουθούν να υπάρχουν προκλήσεις κατά την εφαρμογή τους. Αυτή η πολύπλευρη έρευνα προσφέρει καθοδήγηση σχετικά με τον αντίκτυπο των αποφάσεων διαχείρισης δεδομένων στην αξιολόγηση ημερολογιακών δεδομένων.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Scope of the thesis . . . . .	1
1.2	Patient Reported Outcomes . . . . .	2
1.3	Diary studies . . . . .	4
1.4	Latent variable modelling . . . . .	6
1.4.1	Common Factor Analysis . . . . .	6
1.4.2	Categorical factor analysis . . . . .	12
<b>2</b>	<b>Data handling and modelling options for daily diary data</b>	<b>20</b>
2.1	Use of a portion or aggregation of data . . . . .	22
2.1.1	Single day approach . . . . .	22
2.1.2	Item average approach . . . . .	23
2.2	Use of all the data points . . . . .	24
2.2.1	P-technique . . . . .	24
2.2.2	Design-based models . . . . .	28
2.2.3	Independent analysis . . . . .	29
2.2.4	Multilevel models . . . . .	30
<b>3</b>	<b>Simulation study</b>	<b>40</b>
3.1	Introduction . . . . .	40
3.2	Description of the simulation study . . . . .	42
3.3	Equivalence of graded response model with Confirmatory factor analysis model . . . . .	45
3.4	Methods for exploratory factor analysis . . . . .	46
3.4.1	Modelling methods . . . . .	46
3.4.2	Assessments methods . . . . .	47
3.5	Methods for confirmatory factor analysis . . . . .	50
3.5.1	Modelling methods . . . . .	50
3.5.2	Assessments methods . . . . .	50



<b>4</b>	<b>Exploratory factor analysis</b>	<b>52</b>
4.1	Introduction . . . . .	52
4.2	Descriptive measures for exploratory factor analysis model . . . . .	52
4.2.1	Kaiser Meyer Olkin . . . . .	52
4.2.2	Measure of sampling adequacy . . . . .	53
4.3	Selection criterion for the number of factors . . . . .	54
4.3.1	Kaiser criterion . . . . .	54
4.3.2	Empirical Kaiser criterion . . . . .	54
4.3.3	Parallel analysis . . . . .	55
4.4	Exploratory ordinal factor analysis for single selected say approach . . . . .	56
4.4.1	Input correlation matrix: polychoric correlation . . . . .	56
4.4.2	Model . . . . .	61
4.4.3	Distributional assumptions . . . . .	63
4.4.4	Additional assumptions . . . . .	63
4.4.5	Estimation method . . . . .	64
4.5	Exploratory factor analysis for item average approach . . . . .	66
4.5.1	Input correlation matrix: Pearson correlation matrix . . . . .	66
4.5.2	Model . . . . .	69
4.5.3	Distributional assumptions . . . . .	70
4.5.4	Additional assumptions . . . . .	71
4.5.5	Estimation method . . . . .	71
4.6	Exploratory factor analysis for the split multilevel approach . . . . .	72
4.6.1	Input covariance matrix: Within-individual covariance matrix . . . . .	72
4.6.2	Input covariance matrix: Between-individual covariance matrix . . . . .	72
4.6.3	Model . . . . .	73
4.6.4	Distributional assumptions . . . . .	75
4.6.5	Additional assumptions . . . . .	76
4.6.6	Intraclass Correlation Coefficients . . . . .	76
4.6.7	Estimation method . . . . .	79
4.7	Rotation . . . . .	81
4.8	Goodness of fit measures . . . . .	83
4.8.1	Introduction . . . . .	83
4.8.2	Unscaled fit statistics for Chi-squared, RMSEA, CFI and TLI . . . . .	86
4.8.3	Scaled fit statistics for Chi-squared, RMSEA, CFI and TLI . . . . .	88
4.8.4	SRMR . . . . .	89



<b>5</b>	<b>Confirmatory factor analysis</b>	<b>91</b>
5.1	Introduction . . . . .	91
5.2	Identification of the confirmatory factor analysis model . . . . .	92
5.3	Confirmatory ordinal factor analysis for single selected day approach . . . . .	93
5.3.1	Input correlation matrix: polychoric correlation . . . . .	93
5.3.2	Model . . . . .	93
5.3.3	Distributional assumptions . . . . .	95
5.3.4	Additional assumptions . . . . .	95
5.3.5	Estimation method . . . . .	95
5.4	Confirmatory factor analysis for item average approach . . . . .	95
5.4.1	Input correlation matrix: Pearson correlation matrix . . . . .	95
5.4.2	Model . . . . .	95
5.4.3	Distributional assumptions . . . . .	96
5.4.4	Additional assumptions . . . . .	96
5.4.5	Estimation method . . . . .	97
5.5	Confirmatory multilevel factor analysis . . . . .	97
5.5.1	Input covariance matrix: within and between-individual covariance matrix . . . . .	97
5.5.2	Model . . . . .	97
5.5.3	Distributional assumptions . . . . .	99
5.5.4	Additional assumptions . . . . .	99
5.5.5	Estimation method . . . . .	99
5.6	Goodness of fit measures . . . . .	101
5.7	Heywood cases . . . . .	101
<b>6</b>	<b>Results</b>	<b>103</b>
6.1	EFA results . . . . .	103
6.1.1	Percentage of correct identification of number of factors . . . . .	104
6.1.2	Factor loading strength and range . . . . .	106
6.1.3	ICC results . . . . .	112
6.1.4	Estimated, true and observed inter-item correlation . . . . .	115
6.1.5	Overall factor loading bias and mean squared error . . . . .	119
6.1.6	Goodness of fit measures . . . . .	122
6.1.7	Convergence and Heywood cases . . . . .	127
6.2	CFA results . . . . .	129
6.2.1	Factor loading strength and range . . . . .	130
6.2.2	Estimated, true and observed inter-item correlation . . . . .	135
6.2.3	Overall factor loading bias and mean squared error . . . . .	138



6.2.4	Goodness of fit measures . . . . .	141
6.2.5	Convergence and Heywood cases . . . . .	145
<b>7</b>	<b>Discussion and Conclusions</b>	<b>147</b>
7.1	Discussion on the Results . . . . .	147
7.2	Concluding Remarks . . . . .	152
	<b>References</b>	<b>154</b>
	<b>A Output tables</b>	<b>168</b>
	<b>B Output figures</b>	<b>177</b>
	<b>C Additional IRT models</b>	<b>186</b>



# List of Figures

1.1	Relationship of negative emotion with "Sadness", "Hopelss", and "Irritability" items based on a factor analysis model. . . . .	8
1.2	Relationship of with "Sadness", "Hopeless", "Irritability", "Poor appetite", and "Overeating" items based on a factor analysis model. . . .	10
1.3	Item characteristic curve. . . . .	13
1.4	Item response category characteristic curve of an item with 5 response options. . . . .	14
1.5	Item characteristic curves for a 2-PL model. On the first graph, the difficult parameter is equal to 0 and the slope parameter is equal 1.7. On the second graph, the difficulty parameter is equal to 1 and the slope parameter is equal to 1.7 . . . . .	16
1.6	Item characteristic curves for a 2-PL model. On the first graph, the difficult parameter is equal to 0 and the slope parameter is equal 1.7. On the second graph, the difficulty parameter is equal to 0 and the slope parameter is equal to 1. On the third graph, the difficulty parameter is equal to 0 and the slope parameter is equal to 0.3 . . . . .	17
2.1	Example of P-technique factor analysis model with 2 factors and 6 items.	27
2.2	Single level within-individual model with 2 factors and 4 items . . . . .	36
2.3	Independence model . . . . .	37
2.4	Hypothesized model . . . . .	37
4.1	Bi-variate and marginal latent response distributions for polytomous items $X$ and $Y$ with 6 categories. . . . .	58



6.1 Bar plots for the percentage of iterations (1,000 simulated datasets) in which the factor identification correctly identified the simulated number of factors for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). Single day=Single selected day approach; Within=Within-individual analysis; Between=Between-individual analysis; a=slope parameter from multidimensional graded response model. 105

6.2 Boxplot of the loadings within the first factor across 1,000 simulated datasets for EFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day approach; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 108

6.3 Boxplot of the range of inter item observed and EFA-estimated correlation within the second factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis Single day=Single selected day approach; Item average= Weekly item average; a: Slope parameter from a multidimensional graded response model. . . . . 118

6.4 Overall absolute average bias of the estimated loadings of items within the first factor for the EFA model with 1,000 simulated datasets for 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Item average=Weekly item average; a: Slope parameter from a multidimensional graded response model. . . 120

6.5 Overall MSE of the estimated loadings of items within the first factor for the CFA model with 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Item average=Weekly item average; MSE: Mean square error; a: Slope parameter from a multidimensional graded response model. . . . . 121



6.6 Bar plots for the percentage of poor fit (see Chapter 4.8) in EFA of the 1,000 simulated datasets for each of the data handling approaches for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model. . . . . 124

6.7 Bar plots for the percentage of acceptable fit (see Chapter 4.8) in EFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) based on selected goodness of fit measures. EFA=Exploratory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model. . . . . 125

6.8 Bar plots for the percentage of good fit (see Chapter 4.8) in EFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) based on selected goodness of fit measures. RMSEA=Root mean square error of approximation; EFA=Exploratory factor analysis; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model. . . . . 126

6.9 Boxplot of the loadings within the first factor across 1,000 simulated datasets for CFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Between=between-individual analysis; Within=within-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 131



6.10 Boxplot of the range of inter item observed and CFA-estimated correlation within the first factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day approach; Item average=Weekly item average; a: Slope parameter from a multidimensional graded response model. . . . . 137

6.11 Overall absolute average bias of the estimated loadings of items within the first factor for the CFA model with 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory Factor Analysis; Single day=Single selected day approach; Item average=Weekly item average; a: Slope parameter from a multidimensional graded response model. . . . . 139

6.12 Overall MSE of the estimated loadings of items of within the first factor for the CFA model with 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day approach; Item average=Weekly item average; a: slope parameter. . . . . 140

6.13 Bar plots for the percentage of poor fit (see Chapter 4.8) in CFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model. . . . . 142

6.14 Bar plots for the percentage of acceptable fit (see Chapter 4.8) in CFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; CFA=Confirmatory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model. . . . . 143



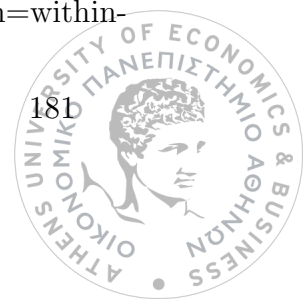
6.15 Bar plots for the percentage of good fit (see Chapter 4.8) in CFA of the 1,000 simulated datasets for each of the data handling approaches for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model. . . . . 144

B.1 Boxplot of the loadings within the second factor across 1,000 simulated datasets for EFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 178

B.2 Boxplot of the loadings within the third factor across 1,000 simulated datasets for EFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 179

B.3 Boxplot of the range of inter item observed and EFA-estimated correlation within the second factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 180

B.4 Boxplot of the range of inter item observed and EFA-estimated correlation within the third factor across 1,000 simulated datasets for scenario 1 and 2 with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 181



B.5 Boxplot of the loadings within the second factor across 1,000 simulated datasets for CFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 182

B.6 Boxplot of the loadings within the third factor across 1,000 simulated datasets for CFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 183

B.7 Boxplot of the range of inter item observed and CFA-estimated correlation within the second factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model. . . . . 184

B.8 Boxplot of the range of inter item observed and CFA-estimated correlation with the third factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis a: Slope parameter from a multidimensional graded response model. . . . . 185



# List of Tables

2.1	Data handling strategies for daily diary data . . . . .	22
3.1	Summary of the literature regarding the within-individual variability in psychological processes and symptoms. . . . .	41
3.2	Scenarios of the simulation study based on the values of slope parameter, and sample size. . . . .	44
3.3	Cut off values for CFI, TLI, RMSEA and SRMR. . . . .	50
4.1	$k_1 \times k_2$ contingency table of $X$ and $Y$ with $k_1, k_2$ categories respectively .	59
6.1	True loadings parameters, and EFA-estimated loadings within factor 1 for the single selected day, weekly item average, within- and between-individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). $a$ : Slope parameter from a multidimensional graded response model. . . . .	109
6.2	Range and average value of the intraclass correlation coefficient for each item across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). $a$ : Slope parameter from a multidimensional graded response model. . . . .	113
6.3	True correlation parameters, observed inter-item correlations, and EFA-estimated correlations for the single selected day and weekly item average approaches with the 1,000 simulated datasets 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). $a$ : Slope parameter from a multidimensional graded response model. . . . .	117



6.4	Percentage of convergence and Heywood case for the single selected day, weekly item average, within- and between-individual analysis approaches for the EFA model with 100, 150, 200, 250 and 350 individuals across 1 week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: slope parameter from a multidimensional graded response model . . . .	128
6.5	True loadings parameters, and CFA-estimated loadings within factor 1 for the single selected day, weekly item average approaches, multilevel CFA with the 1,000 with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2. a: Slope parameter from a multidimensional graded response model. . . . .	132
6.6	True correlation parameters, observed inter-item correlations, and CFA-estimated correlations for the single selected day and weekly item average approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model. . . . .	136
6.7	Percentage of convergence and Heywood case for the single selected day, weekly item average, within- and between-individual analysis approaches for CFA model with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model. . . .	146
A.1	True loadings parameters, and EFA-estimated loadings within the second factor for the single selected day, weekly item average approaches, within- and between- individual analysis approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model. . . . .	169
A.2	True loadings parameters, and EFA-estimated loadings within the third factor for the single selected day, weekly item average approaches, within- and between-individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model. . . . .	171



- A.3 True loadings parameters, and CFA-estimated loadings within the second factor for the single selected day, weekly item average approaches, within- and between- individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model. . . . . 173
- A.4 True loadings parameters, and CFA-estimated loadings within the third factor for the single selected day, weekly item average approaches, within- and between-individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model. . . . . 175



# Chapter 1

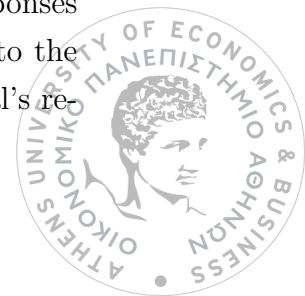
## Introduction

### 1.1 Scope of the thesis

Quite often, Patient Reported Outcome (PRO) data are collected via daily diaries, which involve a collection of a high volume of data across a study period. Such data, can be used for the evaluation of the psychometric properties of a PRO instrument, including the property of the structural validity. This property refers to whether an instrument effectively assesses the variables it claims to evaluate. Such a property could be assessed by the use of a very well-known model in social sciences, the factor analysis model. Traditionally factor analysis has been used in cross-sectional studies (Gorter, Fox, & Twisk, 2015) and as a result the use of single-level factor analysis model is commonly used even when studies involve daily measurements for each individual as in the case of daily diary studies. Apart from this, the popularity of such models is also attributed to the fact that the variability of measurements within individuals (i.e., within-individual variability) is usually considered a statistical nuisance (Schneider & Stone, 2016) rather than an insightful source of variance.

Across the single-level approaches, one can aggregate the data across time by computing the average across a study period for each individual (item average approach) and then use the common factor analysis model. One other approach employs a proportion of the available data by selecting a single day across time (single-selected day approach) and then utilizes categorical factor analysis model.

Both approaches try to use a single data point for each individual for the factor analysis model. However, such models do not capture the multilevel nature of the daily diary data, as observations are nested within each individual. The first level corresponds to the within-individual level, which refers to how individuals responses differ relative to their own typical/average level. The second level corresponds to the between-individual level, which focuses on average differences between individual's re-



sponses. The multilevel nature is not captured in the single-level approaches and as a result they may lead to loss of information, especially when within-individual variability is an insightful source of variance. This is the reason for which an alternative model that could take into account the multilevel nature of the data is warranted, as it could bring additional insights that are not feasible in a single level analysis. Such a model is called multilevel factor model, and it conceptualizes the 2 levels of the data under two different measurement models. However, there have been limited studies of this model in daily diary PRO data as the most commonly used approaches are the item average and the single selected day approach that were mentioned above.

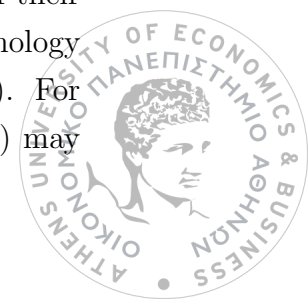
Considering the potential methodological issues that could arise under the commonly used single-level approaches in daily diary PRO data, a simulation study was conducted to investigate such issues. An additional goal was to evaluate whether a multilevel factor analysis model could provide additional insights and what are the possible limitations of such a model.

The current study focuses on factor analysis models under these 3 different approaches, and more specifically this work investigate 2 different frameworks of such models: explanatory and confirmatory framework. In this thesis some basic concepts for PRO, diary studies and latent variable modelling will be introduced. Then the simulation study will be described in detail and a literature review for modelling daily diary data will be presented. Next, factor analysis models for the 3 approaches which are under evaluation in the current simulation study will be described under both the exploratory and confirmatory framework. Finally, the results of the simulation study along with the discussion conclusion will be presented.

## 1.2 Patient Reported Outcomes

PROs are health outcomes that directly come from patients themselves. This means that the insights on the outcome is achieved via direct reporting from patients and usually such an outcome could refer to their functional status, health-related quality of life or symptom burden (van der Willik et al., 2021). The Food and Drug Administration (FDA) defines it as “any report of the status of a patient’s health condition that comes directly from the patient, without interpretation of the patient’s response by a clinician or anyone else” (FDA, 2006).

Patients can express their thoughts and experiences in regard to different aspects of their health related quality of life, or they can assess the symptom severity of their disease. Such insights can be considered as valuable information that the technology or any other observer may not be capable of capturing (Benjamin et al., 2017). For instance, symptoms of fatigue or headache (i.e., symptom frequency or severity) may



not be obvious to observers.

In some occasions, the main goal when studying a disease is the quality of life rather than the survival, and PROs could bring useful insights in such cases. Although outcomes such as survival outcomes, which do not come directly from the patients, enable the investigation of the benefits of a treatment, patient's perspective could bring a more holistic view of the benefits of the treatment of interest (Black, 2013). This is the reason for which patient's perspective is a key when it comes to health case decisions (Baldwin, Spong, Doward, & Gnanasakthy, 2011).

When studying PROs some of the most common terms that are commonly used and that are important to be introduced are:

- PRO instrument/measurement: It is a mean to capture the PRO data
- PRO concept: It is the underlying concept of interest that needs to be measured
- PRO domain: It is a sub-concept which is less broad relative to the PRO concept
- PRO item: It is usually a question or a statement that it is evaluated by patients in order to address a specific concept, such as anxiety
- Endpoint: It is the measurement that will be used to evaluate whether the intervention under investigation is beneficial. Such a measurement should be in compliance with the trial's objectives, design, and data analysis

Generally instruments can be either single or multi-item, which means that the concept of interest can be either measured by one or by multiple items, although one item is not usually capable of adequately measuring complex psychological concepts (Sarstedt & Wilczynski, 2009). An example of a multi-item PRO instrument to illustrate some of the above terminology is the Symptoms of Major Depressive Disorder Scale (SMDDS) (Bushnell et al., 2019) which consists of 16 items. The concept of interest in this case is the symptom severity of major depressive disorder. The PRO domains and the corresponding items of the SMDDS instrument are:

- Negative emotion: Sadness, Hopeless/helpless, Irritability, and Anhedonia
- Anxiety: Feeling overwhelmed and Worry
- Low energy: Tiredness
- Cognition: Intrusive thoughts and Poor concentration
- Sleep Disturbances: General sleep adequacy
- Eating Behaviour: Poor appetite and Overeating



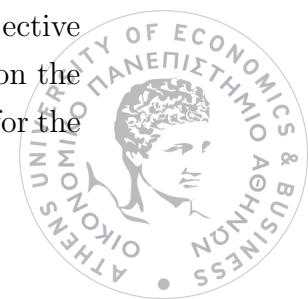
- Low Motivation: Lack of drive, No interest in activities
- Sense of Self: Self-blame: Lack of drive and No interest in activities
- Self Harm/Suicide: Life not worth living

Building upon the understanding of PROs and their terminology, it is important to recognize also the relevance of PRO data in the context of clinical trials. In clinical pharmacology trials, specifically, PRO data could be used either as a primary or supplementary outcome measure (Higgins et al., 2019). They could be used as a primary endpoint in clinical trials where there is not an objective outcome (i.e., quantified via a diagnostic measure) and the outcome of interest can only be retrieved via the subjective perspective of the patient with reference to the impact of the disease. They could be also used as supplementary measures for primary outcomes such as survival rates, and they could provide an additional info regarding patient's symptoms and quality of life (Higgins et al., 2019). For example, they could be a great supplementary measure for providing information in a clinical outcome such as myocardial infraction or acute heart failure (Anker et al., 2014). PROs are useful as they contribute to the evaluation of patients' inclusion or exclusion criteria within a clinical trial, and patients' compliance or non-compliance. For instance, a PRO instrument may provide useful information with regard to medication side effects in patient's quality of life.

### 1.3 Diary studies

Usually, PRO assessments are conducted sporadically, such as when a patient visits a clinician or when attending a clinical visit during a clinical trial. Even so, there are cases when the health outcome of interest is known to fluctuate daily (e.g., mood) and in such cases a high volume of data is required (Bolger, Davis, & Rafaeli, 2003). Such accelerated process of data collection can be implemented via diaries studies.

Examples of such studies are experience sampling (Napa Scollon, Prieto, & Diener, 2009), ambulatory assessment (Smyth & Stone, 2003) and daily diaries studies (Bolger et al., 2003). The goal is to collect data on people repeatedly over time. Experience sampling tries to capture individual self reported experiences via surveys and the main objective is to measure the subjective perspective of patients over time, within the context of everyday life. Such studies are usually conducted in cases where PRO measures are intended for a periodic assessment. Ambulatory assessment focuses on measuring physical states such as heart rate during patient's everyday life, so the main objective is the physiological monitoring (Trull & Ebner-Priemer, 2014). In daily diaries, on the other hand under interval-contingent sampling (i.e., predefined fixed time-points for the

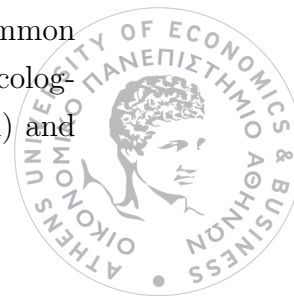


data collection), the assessments are collected daily (once a day). One main difference between such studies with experience sampling is that for the latter type of study the data collection could occur across multiple times within the same day, while for the former type of study the assessments usually occur once a day (Horstmann, 2021). Additionally, daily diaries could also be event-based, where patients provide their self reports after the event of interest occurs (Lischetzke & Könen, 2020).

In daily diaries studies, which will be the scope of this thesis, participants provide frequent reports of their aspects of their daily life based on self report questionnaires. These reports are usually in the form of surveys, so they are comprised of multiple items/questions in order to study an underlying phenomenon, such as for example anxiety or depression. The two main aims of studying such a phenomenon are the investigation of day to day individual's fluctuation across time and the comparison of different groups (Van de Schoot, Lugtig, & Hox, 2012). In health related studies more specifically, the items could be used to measure a variety of concepts regarding the quality of patient's health life or patient's symptom severity of a disease. This is accomplished by measuring the day-to-day fluctuation of subject's health status over time (Gorter et al., 2015).

Although such an insight is also retrieved in the traditional longitudinal designs, where each participant provides data only a few times over a large time interval, daily diary studies provide a high volume of data per person, which enables the decomposition of the observed variability into two different levels (i.e., within- and between-individual levels). As a consequence, this dynamic process of data collection in such studies gives a large room to investigate research questions with reference to the within- and between-individual variability (Hamaker & Wichers, 2017; Ram & Gerstorf, 2009). The within-individual variability tries to measure how an individual's response variate relative to his/her own average/typical level, and between-individual variability tries to measure how individual's responses varies on average. Although within-individual variability is often considered as a statistical nuisance or measurement error (Schneider & Stone, 2016), studies have suggested that when it comes to individual symptoms, useful information could be retrieved based on the within-individual source of variance (Ram & Gerstorf, 2009). So daily diary data are an important modality for capturing day to day fluctuation if a researcher is interested in the within-individual variability.

A reason for the development and the popularity of daily diary studies is the technological development which has enabled people to measure different aspects of their health status in higher frequency across short time intervals. So the collection of real time data through electronic devices such as phones or smartwatches became a common and convenient data collection method. These studies also led to the increase of ecological validity (i.e., the ability of the generalization of the results to the real-world) and



the reduction of patient burden as patient's responses are captured in real-time, so there is not any need for patient to recall any information or experience (de Haan-Rietdijk, Voelkle, Keijsers, & Hamaker, 2017).

The operationalization of such studies continues to grow as the above advantages have gained some recognition in the area of PRO (Schneider & Stone, 2016). The reduction of recall bias and the investigation of day to day variability in patient's symptoms have made daily diary data an important modality for PROs. That is because they can contribute additional value compared to other type of studies, such as cross-sectional studies where there is available information about the individual at only one single timepoint, so there is no available information regarding the individual's variability across a study period.

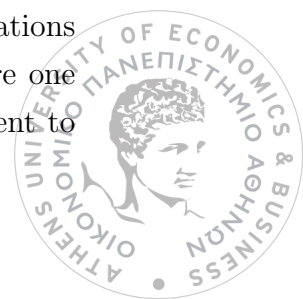
In the case of cross-sectional studies, the individual response is compared to the general patterns of the group rather than his/her own fluctuations across time. So the main interest in cross-sectional studies is how an individual's responses change in comparison with the general pattern of change across all individuals. An assumption cross-sectional studies make is that individual state remains stable across time, but that is not always the case as individual behaviour could be significantly affected by situational conditions, such as in the case of the mood (Moskowitz & Young, 2006). It has been empirically showed that concepts such as mood or personality might show a great fluctuation across time (Hooker, 1991) and they might not be as stable as they were previously hypothesized (Hamaker & Wichers, 2017). Psychological processes could impact both the level of the outcome and the variation of that outcome within groups and this is the reason for which daily diary studies are useful as they could capture both sources of information (Bryk & Raudenbush, 1988).

Overall, daily diary studies are important as they could capture patient's thoughts and experiences regarding different aspects of their health-related-quality of life. The applicability of such studies is apparent as they are more representative of the complex human behaviour in everyday life, which could variate across days or across different time intervals of the same day. This is the reason for which the investigation of PRO in such studies could be valuable.

## 1.4 Latent variable modelling

### 1.4.1 Common Factor Analysis

An important goal in the PRO analysis studies is the investigation of items associations within a questionnaire. Questionnaire items are intended to collectively measure one or more target constructs, so the pattern of association between items is pertinent to



understand the expected homogeneity within a measure and the reliability and validity of the measure. PRO questionnaires usually include items related to the symptoms of a disease, and they measure one underlying construct (i.e., overall symptom severity). This is the reason for which the concept of interest that a PRO instrument tries to capture may be considered as a latent variable. Thus, it is important to capture patient's feedback by examining the relationship of the items of a questionnaire that they were administered across a study period.

For instance, in the case of the SMDDS instrument, questions such as: "Feeling overwhelmed" and "Worry" try to reflect anxiety which is an unmeasured/unobserved concept. Such concepts are usually called factors, as they are comprised of multiple items that measure the same concept of interest. That is to say, the reason the "Feeling overwhelmed" and "Worry" are related is attributed to the fact that they try to reflect the same unobserved factor. This necessitates the construction of a measurement model that could describe the relation between the observed (items) and latent variables (factors). This is the reason for which latent variable models were developed to conceptualize that type of relationship through a measurement model (Bishop, 1998).

Their utility is apparent in psychology and mental health studies as usually in such fields the concepts of interest are abstract such as pain, depression, or anxiety levels. Such concepts are difficult to be quantified as there are many components that pose an impact on them. Although it is difficult to find a single characteristic to measure how severe are the symptoms of depression, it is possible to capture a variety of aspects of that mental illness that are measurable, such as the hours of sleep, the level of motivation or concentration. One of the most well known models that try to explain the relationship among those observed characteristics is factor analysis (Spearman, 1961). Particularly, assessments for measuring the patient's health status are built on the assumption that responses to items/questions on the PRO instrument are informed by latent constructs that are the intended target of the measurement. This is the reason for which models such as factor analysis models are appropriate for such data. More specifically, factor analysis models aim to quantify, comprehend and explain the relationship of responses to PRO instruments with the underlying health-related issue of interest (Fayers & Machin, 2013).

In factor analysis models, one should consider 3 important components for its construction:

- The observed or manifest variables which are usually items/questions of a questionnaire
- The unobserved variables which are called factors
- The correlation between items



The third component, which is the correlation between items, is usually used as an input for fitting the factor analysis model. Generally, it is expected that items that explain the same concept of interest will be highly correlated and items that measure a different concept will be lowly correlated. Figure 1.1 illustrates an example of how the correlation of negative emotion is described with the corresponding observed variables/items of the SMDDS instrument based on a factor analysis model. The unidirectional arrow from depression to the "Sadness", "Hopeless", and "Irritability" items depict their association strength based on parameters  $\lambda_{11}$ ,  $\lambda_{21}$ , and  $\lambda_{31}$  respectively, which are commonly called loadings. On the other hand,  $\psi_1$ ,  $\psi_2$  and  $\psi_3$  are error variances or unique variances and could be considered as the amount of variances of the items that is not explained based on their relationship with the negative emotion. If the error is high, that implies that there might be other reasons apart from the latent construct that explain the relation of the items. Finally,  $\phi_1$  is the variance of the negative emotion.

Figure 1.1 is a called path diagram. Such a diagram consists of boxes and circles which are connected by arrows. Observed variables are represented by square boxes and latent factors are represented by circles. Single headed arrows define the relationship between the observed variables and the latent factors, with the variable at the tail of the arrow causing the variable at the point. This relationship is usually quantified via a regression coefficient.

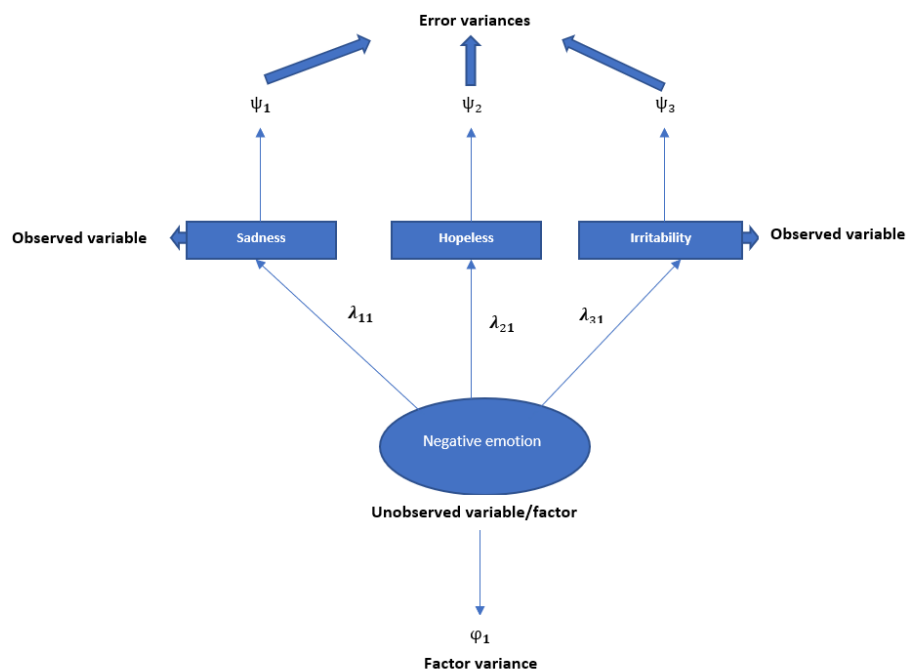


Figure 1.1: Relationship of negative emotion with "Sadness", "Hopelss", and "Irritability" items based on a factor analysis model.

The measurement model of Figure 1.1 is also described in equation (1.1)

$$y_i = \lambda_{i1}\eta_1 + \epsilon_i \quad (1.1)$$

where:

$y_1$  : Value of the "Sadness" item

$y_2$  : Value of the "Hopeless" item

$y_3$  : Value of "Irritability" item

$\lambda_{11}$  : Association strength between negative emotion and "Sadness" item

$\lambda_{21}$  : Association strength between negative emotion and "Hopeless" item

$\lambda_{31}$  : Association strength between negative emotion and "Irritability" item

$\epsilon_1$  : Residual for the "Sadness" item

$\epsilon_2$  : Residual for the "Hopeless" item

$\epsilon_3$  : Residual for "Irritability" item

$\eta_1$  : The factor score of the negative emotion

Items could be seen as endogenous variables as they are determined based on their relationship with other variables, whereas the factors as exogenous as they are not caused by other variables in this example.

Usually when studying a psychological process there is not 1 factor present, as in the example of SMDDS instrument. Along with negative emotion, there are other factors such as eating behaviour. In many cases such factors could be related and this kind of information is also captured in factor analysis models. PRO instruments generally could capture either unidimensional (1 factor) such as overall symptom severity or multidimensional (multiple factors) concepts such as health related quality of life.

On Figure 1.2 we can see how the relationship between items is expressed when more than 1 factor is present based on a factor analysis model. An additional parameter that is introduced is,  $\phi_{12}$  which expresses the correlation between factors. Note that, typically, it is recommended to have at least 3 items per factor when conducting factor analysis. So this example is provided solely for illustrative purposes.



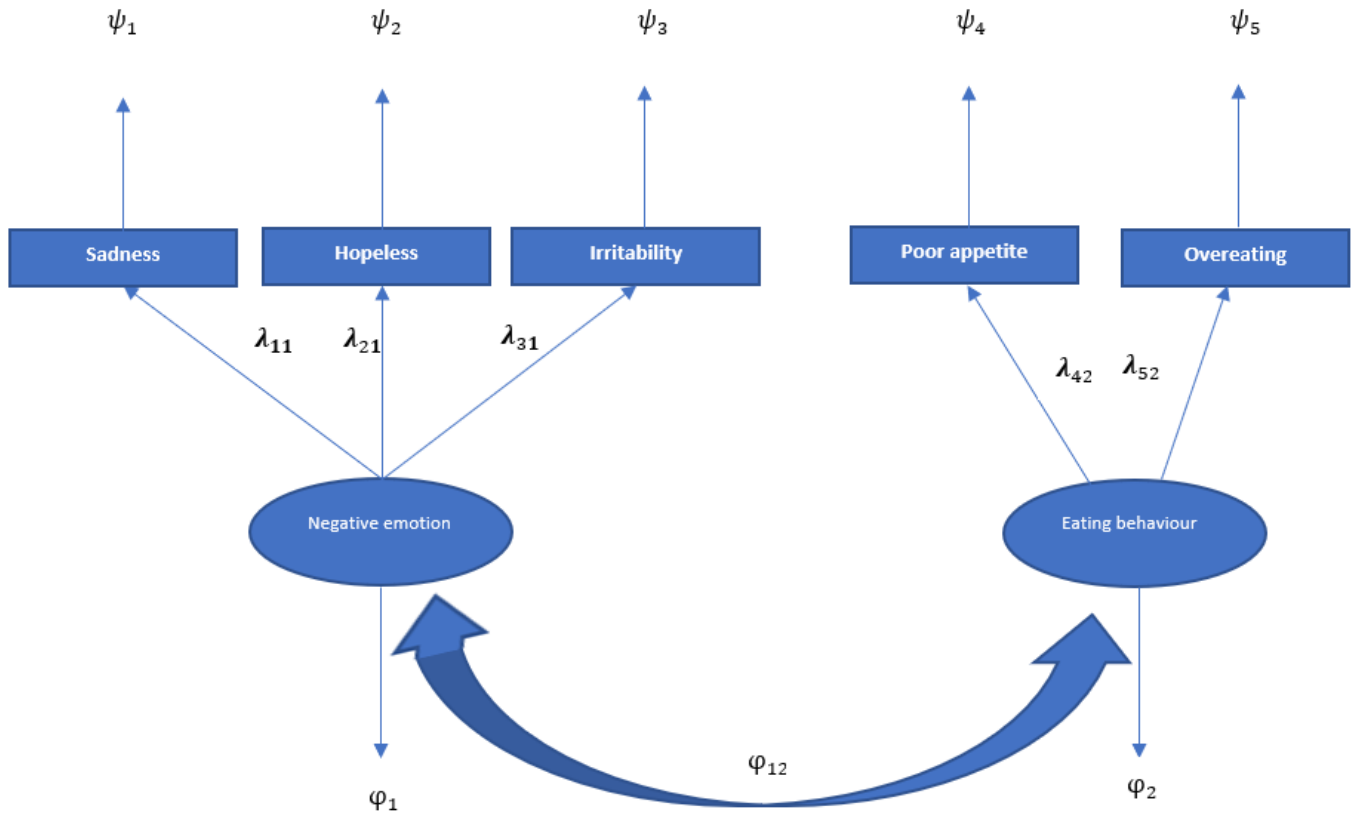


Figure 1.2: Relationship of with “Sadness”, “Hopeless”, “Irritability”, “Poor appetite”, and “Overeating” items based on a factor analysis model.

The measurement model of Figure 1.2 is also described in equation 1.2 and 1.3

$$y_i = \lambda_{ij}\eta_j + \epsilon_i \quad (1.2)$$

where:

$y_i$  : Item  $i$

$\lambda_{ij}$  : Loading of item  $i$  on factor  $j$

$\epsilon_i$  : Residual of the  $i$  item

$\eta_j$  : Factor score for the  $j$  factor

Alternatively the model could be written in a different and a more natural way in a sense that factor analysis models try to describe the relationship between the items,

so it tries to estimate the covariance matrix. The alternative form is described in 1.3.

$$\Sigma = \Lambda Cov(\eta)\Lambda' + \Psi, \quad (1.3)$$

where

$$Cov(\eta) = \begin{pmatrix} \phi_1 & \phi_{12} \\ \phi_{21} & \phi_2 \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{53} \end{pmatrix}, \Psi = \begin{pmatrix} \psi_1 & 0 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 \\ 0 & 0 & 0 & \psi_4 & 0 \\ 0 & 0 & 0 & 0 & \psi_5 \end{pmatrix}$$

$Cov(\eta)$  : 2x2 Covariance matrix of factors

$\phi_{ij}$  : Covariance between  $i$  and  $j$  factor

$\phi_i$  : Variance of factor  $i$

$\Lambda$  : 5x3 Loading matrix

$\lambda_{ij}$  : Loading of item  $i$  on factor  $j$

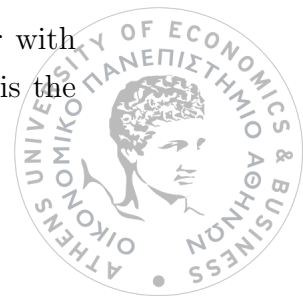
$\Psi$  : 5x5 matrix with the error variances

$\psi_i$  : Error variance for the  $i$  item

Note that in the example above, the relationship of items is already known. This is the reason for which the loadings were assumed to be 0 for some of the items, as they were related with only one factor. Although this was true for this example, quite often the relationship of the items and factors is not known, so in that case it is necessary to estimate the association strength of each item with all the factors. The first type of model which was illustrated in this example is called Confirmatory factor analysis (CFA) and the second type of model is called Exploratory factor analysis model (EFA). In EFA there is not an a priori knowledge of the number of factors or the number of indicators per factor, while in CFA such information is known based on previous studies or a theoretical scientific justification.

Both type of models share the following assumptions:

- $E(\eta)=0$  where  $\eta$  is a  $k$ -dimensional vector with the factors
- $Cov(\eta) = I$  where  $I$  is a  $k \times k$  identity matrix,  $\eta$  is a  $k$ -dimensional vector with the factor,  $Cov(\eta)$  is a  $k \times k$  covariance matrix between the factors, and  $k$  is the number of factors



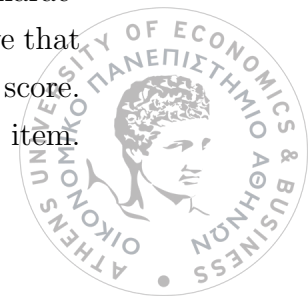
- $E(\epsilon)=0$   $\epsilon$  is a  $p$ -dimensional vector with the residuals, and  $p$  is the number of variables
- $Cov(\epsilon)=\Psi$  where  $Cov(\epsilon)$  is  $p \times p$  covariance matrix of the residuals, and  $p$  is the number of variables (this matrix is a diagonal matrix as in the above example)
- $Cov(\epsilon_i, \eta_j)=0$  for every  $i \neq j$  where  $\epsilon_i$  is the residual for  $i$  variable and  $\eta_j$  is the  $j$  factor
- The observed data comes from multivariate normality

### 1.4.2 Categorical factor analysis

The development of factor analysis models and their methodology were developed for application with continuous data. However, in health related studies, the items of a questionnaire are usually measured on an ordinal or binary scale. So the methodology for these type of data should be adjusted accordingly. This means that a measurement model is needed to quantify the relation between the observed categorical item responses and the continuous latent variables. This is the reason for which categorical factor analysis (Thissen & Steinberg, 1986) was introduced to fill this gap which is employed for the proper analysis of ordered and binary data. The idea is that the categorical responses to questionnaire items are a manifestation of some latent continuous variables (Wirth & Edwards, 2007). The response options thus reflect the categorization of this latent variable based on proposed thresholds. These thresholds could be seen as the quantiles of the distribution of these latent continuous variables.

These models were developed based on a family of models which is called Item Response models (Thissen & Steinberg, 1986). The categorical factor analysis models were developed with the aid of Item Response Theory (IRT) framework. This is the reason for which IRT basic concepts will be introduced as a basis for understanding the categorical factor analysis. As in the case of factor analysis models, where the goal is to explain the relationship among items with the latent construct through the correlation of items, IRT model tries also to explain this relationship but with a different perspective. It tries to explain the relationship among the individual items with the latent construct being measured. More specifically, they describe how individual ability, which is typically conceptualised as a latent trait, is related with an individual response to an item of the questionnaire.

This relationship is usually visualized through a graph, which is called Item characteristic curve. An example of this is illustrated in Figure 1.3 where there is a curve that describes how the probability of endorsing an item is affected by the latent trait score. The higher the latent score is, the higher the individual ability to endorse the item.



Although this example applies for binary data, the same logic could be adopted for polytomous items as well. In that case, there will be a curve for each response category, explaining how the probability of endorsing each category is affected by the individual's latent score. This curve is called item response categorical characteristic curve, and it could be considered as the item characteristic curve for the case of polytomous items. An example of this is illustrated in Figure 1.4

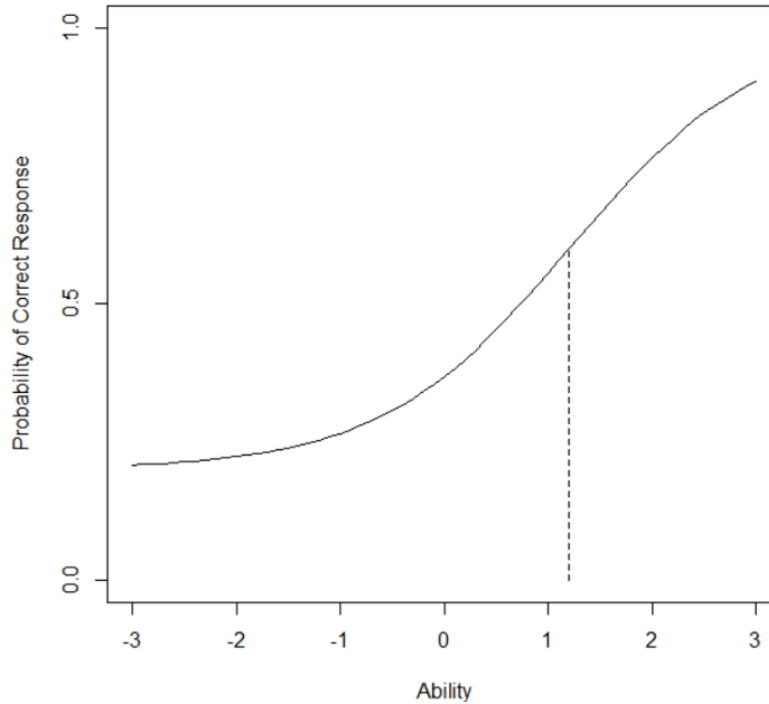


Figure 1.3: Item characteristic curve.

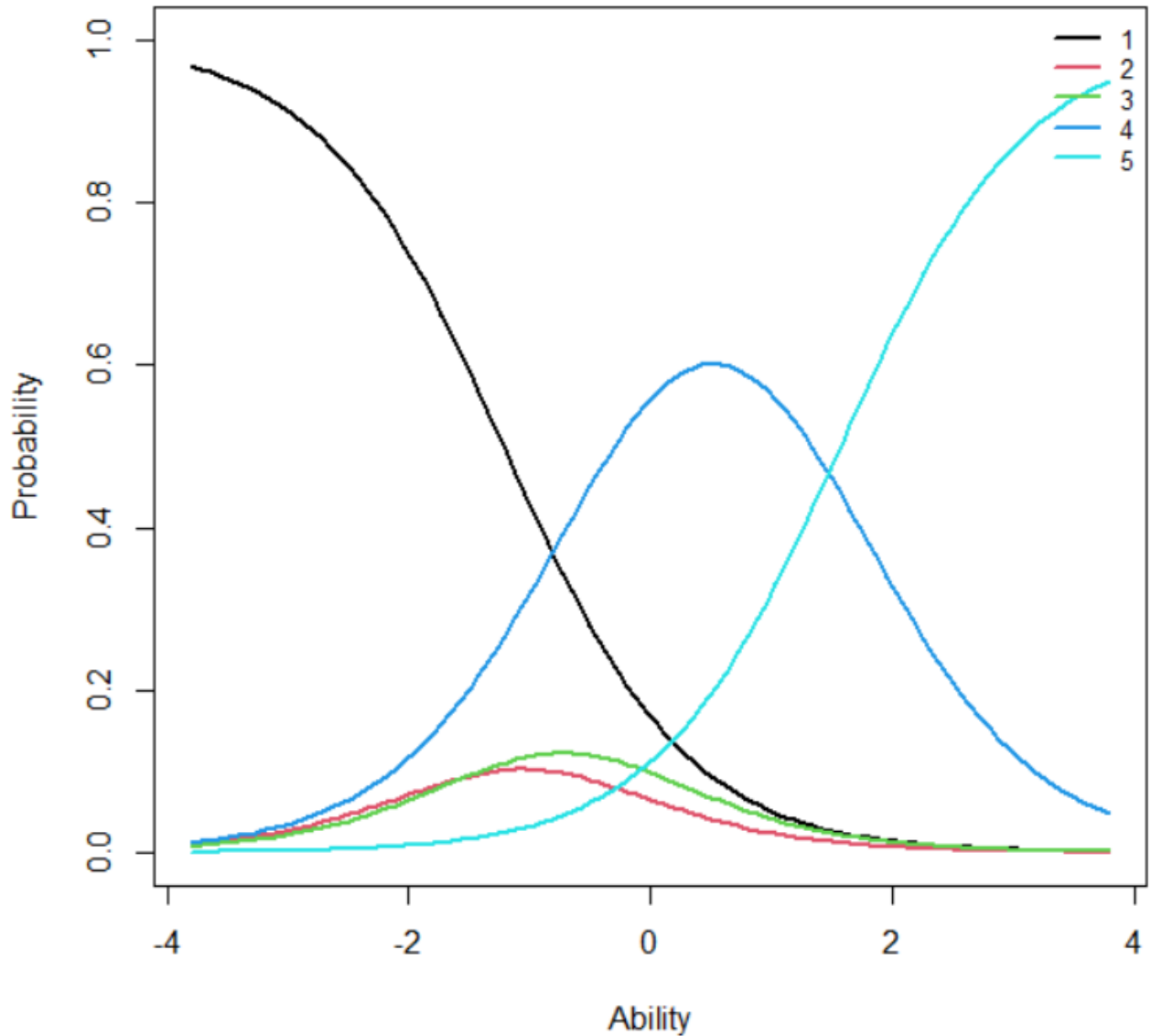


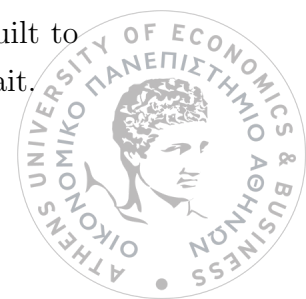
Figure 1.4: Item response category characteristic curve of an item with 5 response options.

The basic parameters that are needed for the construction of item response models are:

- Difficulty parameter ( $b$ )
- Slope parameter ( $a$ )
- Theta ( $\theta$ )
- Guessing parameter ( $c$ )
- Scaling factor ( $D$ )

Difficulty parameter, which could be also called location or threshold parameter, is the value of the latent trait for which the item functions best. More specifically, is the value of the latent trait for which the probability of endorsing an item is 50%. For polytomous items, this could be interpreted as the point of the latent trait, where there is separation between 2 consecutive response categories. Slope parameter, which is also called discrimination parameter quantifies how an item can discriminate between subjects with different levels of latent trait. Theta is the score of the latent trait, which quantifies individual's 'ability'. Guessing parameter describes whether the probability of responding to the item is attributed to guessing or not. It could be described as the probability of an individual with a low latent trait score or ability to answer an item correctly.  $c$  has a theoretical range from 0 to 1, but in practice values above 0.35 are not considered acceptable (Baker, 2001). When studying the quality of life or physical functioning, such a parameter has no application, and it is usually useful in an educational setting where answering a question randomly is possible. Finally,  $D$  ranges from 1 to 1.7, and it usually set equal to 1.7. Its use is to bring the logistic metric estimates close to the normal ogive model (Reckase, 2009). Although at first normal ogive function or alternatively probit function was used based on the cumulative normal distribution (Baker, 2001; Birnbaum, Lord, & Novick, 1968), there were many mathematical challenges. This is the reason for which logistic models were utilized, as their implementation was easier.

To illustrate an example of the interpretation of the difficulty and slope parameters, Figure 1.5 and 1.6 are illustrated below. These figures illustrate the impact of modifying the values of the parameters on the item characteristic curve and the performance of the questionnaire. Figure 1.5 illustrates that when  $a$  is constant and  $b$  shows an increase, an individual needs to have a higher score in order to have a probability of 50% endorsing the correct response. This means that this parameter could provide information on whether an item is considered "difficult" or "easy". If an item is easy, then even someone with a low latent score could have at least 50% of endorsing the item. On the other hand, if the difficulty parameter stays constant and the slope parameters decrease, the probability of endorsing an item is starting to become constant. What this means is that even when someone has higher value of latent score which imply higher "ability" or "symptom severity" in the context of PRO data, the probability of endorsing an item does not seem to change. This is the reason for which in Figure 1.6 the first graph is an ideal graph as those who have low latent score compared with those who have high latent trait have different probabilities of endorsing the item. This information could be useful to inform about whether an item is appropriately built to discriminate an individual's ability with theoretically different levels of latent trait.



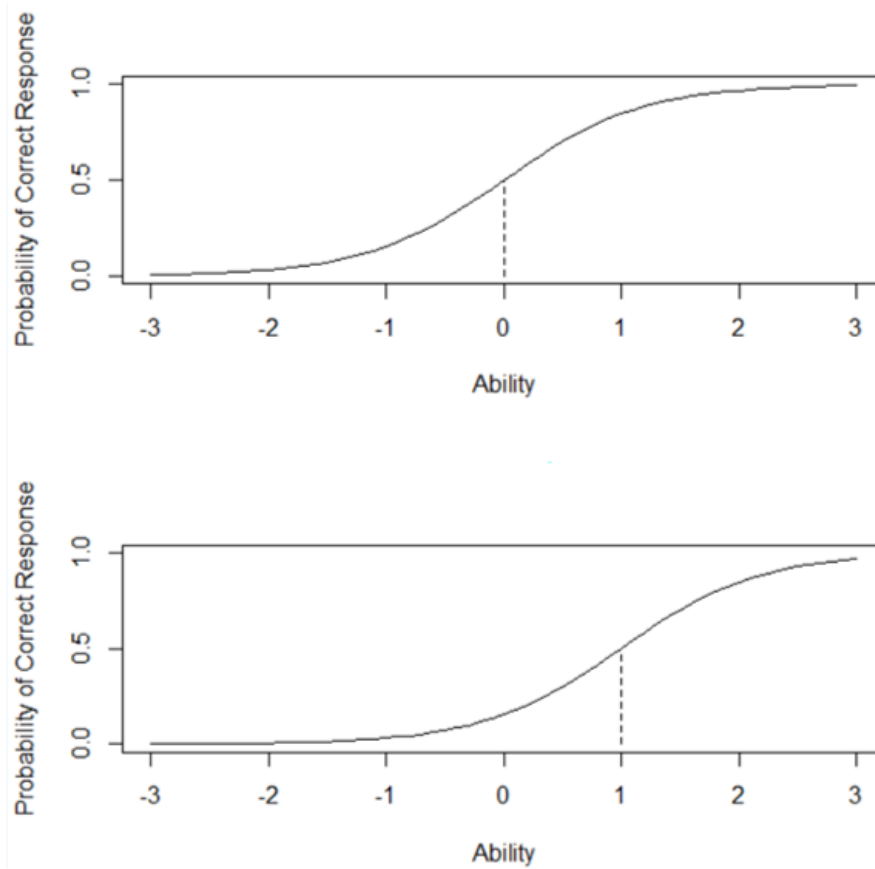


Figure 1.5: Item characteristic curves for a 2-PL model. On the first graph, the difficult parameter is equal to 0 and the slope parameter is equal 1.7. On the second graph, the difficulty parameter is equal to 1 and the slope parameter is equal to 1.7

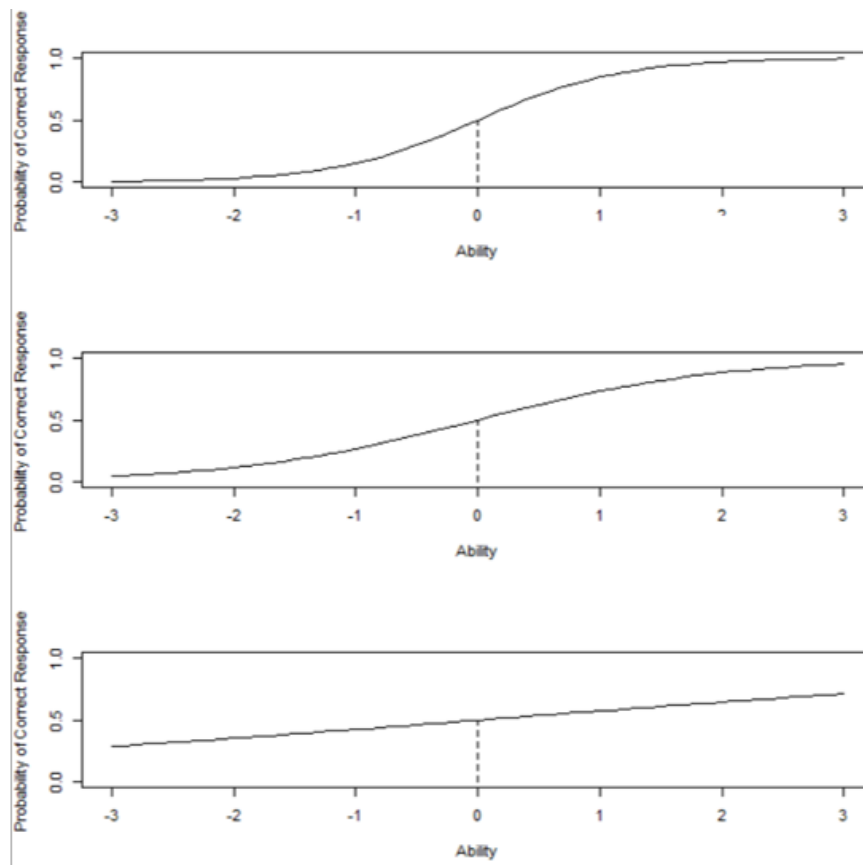


Figure 1.6: Item characteristic curves for a 2-PL model. On the first graph, the difficult parameter is equal to 0 and the slope parameter is equal 1.7. On the second graph, the difficulty parameter is equal to 0 and the slope parameter is equal to 1. On the third graph, the difficulty parameter is equal to 0 and the slope parameter is equal to 0.3

The IRT models are quite useful when it comes to the development of a PRO instrument (Nguyen, Han, Kim, & Chan, 2014) as they could provide a very detail description of the item performance of the questionnaire of interest. There is a wide range of such models, giving the opportunity to accommodate for different measurement situations. Some of the criteria for deciding which model to use are:

- Whether the item scale is polytomous or binary
- Whether the discrimination parameter can be assumed to be constant across all items
- Whether guessing parameter could have any relevance on the questionnaire of interest
- Whether category response parameters should be kept constant across the item

Although there are many models under the umbrella of the IRT models such as Rasch model, 2 parameter logistic and 3 parameter logistic models (see Appendix C),

a more flexible model will be presented here as it is used as the basis for producing PRO data, which is called graded response model. Such a model is also equivalent with a model under evaluation for this work (i.e., ordinal factor analysis) (Takane & De Leeuw, 1987).

This model could be considered as a generalization of the 2-PL model (Wirth & Edwards, 2007) as their difference is that the graded response model assumes that the difficulty parameter is different of each item level. The graded response model is described in equation 1.4 and 1.5

$$P(Y_{ik} = y_{ik} | \theta_k) = P_{y_{ik}}^*(\theta_k) - P_{y_{ik}+1}^*(\theta_k) \quad (1.4)$$

where:

$$P_{y_{ik}}^*(\theta_k) = P(Y_{ik} \geq y_{ik}) = \frac{1}{1 + \exp(-Da_i(\theta_k - b_{ij}))} \quad (1.5)$$

where:

$\theta_k$  : Latent trait for  $k$  individual

$b_{ij}$  : Difficulty parameter for item  $i$  and item level  $j$

$D$  : The scaling factor

$a_i$  : Slope parameter for the item  $i$

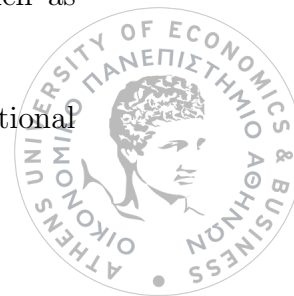
$Y_{ik}$  : Response of individual  $k$  on item  $i$

$P_{y_{ij}}^*(\theta_k)$  is called boundary function, and it represents the probability of responding an item response greater or equal than level  $K$ . If item categories are  $K+1$  then the boundary functions are  $K$ .

Although these models are important for understanding the relationship of individual's response to a questionnaire with individual's latent trait/ability, there are some assumptions that someone has to have in mind when proceeding to analysing such models.

The key assumptions are:

- Unidimensionality of the measured concept: The number of dimensions/factors for the concept of interest is one. Although this assumption has to be met for the above models, there have been developments of multidimensional models where the number of dimensions could be assumed to be greater than one, such as multidimensional graded response model (Muraki & Carlson, 1995)
- Local independence: Items are only related due to the latent trait, so conditional



to latent trait the items are independent

- Monotonicity: As the individual's latent trait/ability increases the probability of endorsing the item will also increase
- Item invariance: Items remain constant across different population groups

IRT models, such as graded response model, are quite useful as they are flexible. This flexibility could be also employed when someone is interested in simulating ordinal data, especially when trying to simulate PRO data as these model as usually used for fitting such data. Although it is common that these type of models assume unidimensionality of the latent structure, this work tries to utilize a more flexible model which assumes multidimensionality of the factor structure by using multidimensional graded response model. This model will be used for the simulation study of this thesis, which will be described in the Chapter 3.



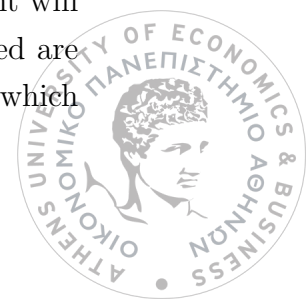
## Chapter 2

# Data handling and modelling options for daily diary data

There is a plethora of data handling approaches and modelling options when it comes to the assessment of structural properties of a daily diary data instrument. This is attributed to the multilevel nature of the data, as observations are nested within each individual. Traditionally in such data where there is intra-individual nested structure, aggregation, desegregation of the data into one single level or selecting a proportion of data (e.g., select a single day) are usually employed to avoid the challenges that are inherent due to the dependencies of the data within individuals.

Traditionally, data aggregation, or selection of a proportion of data (e.g., selection of one single day) are usually employed in daily diary PRO evaluation (Stone, Broderick, & Kaell, 2010; Stone, Broderick, Schneider, & Schwartz, 2012) to handle the inherited challenges (i.e., interdependencies of data within the individuals). On the contrary, desegregation method, which refers to analysing the data as if they were independent, is not usually employed in daily diary PRO data as it may lead to wrong inference. However, it will be introduced for contextualisation.

Factor analysis had been traditionally used in cross-sectional studies (Gorter et al., 2015), where a single time point for each individual is used to represent the individual's profile. Even in the case of daily diary studies, where there are repeated measurements across individuals, single level analysis is usually being employed. The reason behind this is the fact that within-individual variability is often considered a statistical nuisance (Schneider & Stone, 2016). This is the reason for which usually items are averaged across time in order to reduce the noise or a single day is selected by assuming that all days will provide the same insights given that the construct of the instrument will remain constant. However, many psychological processes as previously mentioned are not as stable as initially hypothesized. This means that the assumptions under which



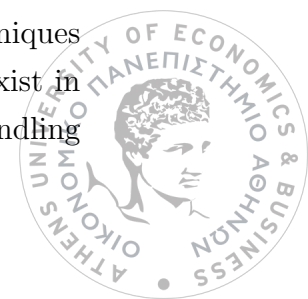
such simplistic approaches are utilized are not always true. So further consideration is required on whether averaging the item scores across time or selecting a single day are appropriate in terms of their insights they provide, or they lead to loss of information as within-individual variability is not taken into account.

This led to the notion that day-to-day fluctuations of individuals could be insightful and a key element for understanding individual change. It could be even studied as a distinct construct itself (Ram & Gerstorf, 2009). Such a construct is investigated by studying the association of the deviation of individual responses and the individual's average/typical level across various items. In simple terms, this approach enables researchers to study how individuals' responses fluctuate around their own personal typical level, which provides insights into the inherent variability within individuals. Such a consideration necessitates a model that could allow researchers to investigate the construct of patient's change into multiple levels (i.e., within- and between-individual level). Models serving this purpose are usually called multilevel models, and they will be the main focus of this Chapter.

The commonly adopted strategies that do not take into account the dependency of the data and the limited use of multilevel models in daily diary PRO data create a large room for investigation. The nature and the complexity of these data requires a careful consideration regarding the data handling and modelling procedure that need to be followed. This implies that potential methodological issues could arise in the simplistic approaches when within-individual variability is not considered a statistical nuisance, which is an important issue that warrants investigation. The intention of this work is to better understand what methodological issues could arise in well-established approaches when doing single level analysis, as to inform the advantages and disadvantages of the methodologies. This work also aims to seek what additional insights could be retrieved if one is interested to study the factor structure of the data into two separate levels via multilevel models and investigate their potential limitations.

As previously mentioned, the multilevel nature of the data leads to numerous data handling options, and each of them serves a different purpose. Although the methods under investigation are single selected day, item average approach and multilevel models, additional methods will be also introduced for educational purposes and to provide a more spherical overview of the potential strategies used to analyse daily diary data. Although these additional strategies are not typically used when evaluating a dairy PRO instrument, they are presented as they could theoretically be implemented due to the multilevel nature of the data.

In this Chapter, a variety of data handling options and factor analysis techniques will be presented in order to provide an overview of different approaches that exist in the area of factor analysis with applicability to daily diary data. For the data handling



options that utilize a proportion or aggregation of the data, single selected day approach and item average approach will be presented. Then data handling options using all data such P-technique, design-based approach, and independent analysis will be briefly presented. Multilevel factor analysis models will then follow. Table 2.1 provide a brief overview of the different data handling approaches that will be presented in this Chapter.

Table 2.1: Data handling strategies for daily diary data

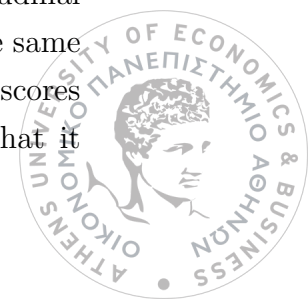
<b>Data handling strategy</b>	<b>Name of the strategy</b>	<b>Purpose</b>
Data aggregation: average the item score across time for each individual.	Item average approach.	Study the item average correlation.
Use of a proportion of the data: select a single day across time.	Single selected day approach.	Study individual item correlation.
Use of all data points.	Design-based approach.	Make valid inference compared to the independent analysis by creating an overall model (single-level).
	P-technique.	Study within-individual variability.
	Independent analysis.	Take a first overview of the latent structure of the data.
	Multilevel modelling.	Take into account the multilevel nature of the data by creating a measurement model for each level.

## 2.1 Use of a portion or aggregation of data

In this Section, single selected day and item average approach will be described.

### 2.1.1 Single day approach

Among the simplistic approaches for analysing the latent structure of a diary instrument is to make an evaluation for a single day (e.g. the first, final or a random day of seven days across a week). This approach is helpful when the main goal is to study the correlation of individual items, and it is implemented under the assumption that scalar invariance holds across time. Assuming that the instrument is intended for longitudinal measurement, this means that the construct is being measured across time in the same way. After selecting a single day, factor analysis is then performed on the item scores from that single timepoint. One of the main advantages of this method is that it



sidesteps challenges around dependencies at the within-individual level (Reise, Ventura, Nuechterlein, & Kim, 2005). This is the reason for which it is a commonly adopted strategy in diary-based PRO instrument validation (Gater et al., 2022; Lipton et al., 2022; Martin Nguyen, Bacci, Dicipinigaitis, & Vernon, 2020).

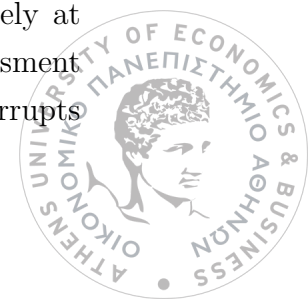
However, there are methodological issues to this approach that warrant acknowledgement and consideration (whether the day is randomly selected or, for example, selected on the basis of an assessment schedule such as 7 days post-baseline). This method does not account for all the sources of variances in the data, and this can lead to biased estimations (J. Little, 2013) and inflated estimates of variability (Stone et al., 2010). The assumption that the selected day is adequately representative of alternative timepoints may not always hold.

### 2.1.2 Item average approach

Another approach that is traditionally employed is uni-variate summary of the response profile for each individual respondent, in which the repeated measures data are aggregated or combined through a summary statistic (Diggle et al., 2002; Omar, Wright, Turner, & Thompson, 1999). A common summary method is to use the average (mean) of observations across time for each individual (Matthews, Altman, Campbell, & Royston, 1990; Frison & Pocock, 1992). The latter is useful when the main target of investigation is the differences at the between-individual level (J. Hox, 1998) and when the correlation at the averaged item level is of interest. This approach is useful for example when someone is interested to study average levels of individual's pain (Broderick, Schwartz, Schneider, & Stone, 2009).

This strategy is also useful when investigating PRO assessments of the same individuals at various timepoints (Stone et al., 2012). The benefits of computing the average of item scores across time lie partly in the simplicity of its implementation. Another advantage is that the summary statistics can be potentially calculated even if there are some missing data or the number of observations differs among subjects (Matthews et al., 1990).

When missing values occur non randomly, as can arise with assessments of health-related quality of life during clinical trials (Fairclough, Peterson, & Chang, 1998), using a derived variable approach such as single averages for each patient may not be an efficient data handling method (Griffiths, Williams, & Brohan, 2022). Nonetheless, there is no universal method for handling such missing data properly (Miettinen, 2012; Vach, 2012; R. J. Little, 1992). Only when missing data are missing completely at random, such as when a patient forgets to complete the questionnaire (in the assessment of conditions where recollection difficulties are not a symptom) or the device corrupts



the data, would simplistic approaches such as the averaging method yield unbiased estimations (Greenland & Finkle, 1995).

There are several further limitations to the item average approach. For example, as this method ignores within-individual variability across time (Reise et al., 2005), bias may occur in the estimated standard errors. The correlation of observations on the within-individual level could lead to reduced estimates on that level when compared with independent data that lack such interdependencies (Bolger et al., 2003).

Consequently, while the weekly item average resolves many of the practical data handling challenges, this approach may also have unintended consequences. Data handling strategies that model rather than eliminate the variability inherent in daily diary data should be considered.

## 2.2 Use of all the data points

### 2.2.1 P-technique

When someone tries to explore intra individual change across time, a very widely used method is P-technique (Cattell, 1963). Such a method has been used in daily diary studies (Brose & Ram, 2012; Kurz, Johnson, Kellum, & Wilson, 2019; Foster & Beltz, 2021) as it involves a high volume of data for each individual, which enables the use of time series models. P-technique focuses on detecting a pattern of systematic change for each study participant and then comparing it across individuals to evaluate the relative peculiarity or the generalization of the changing pattern (Jones & Nesselroade, 1990).

The main goal of this method is to obtain the individual's scores from a multi item questionnaire of a latent construct across multiple time points (I. A. Lee & Little, 2012). This is illustrated by factor analysing the within-individual covariance for each individual separately (Reise et al., 2005). The results are drawn for a single individual and in order to generalize the results factor invariance must be tested, which refers to the equality of loadings across individuals. This is tested through the coefficient of congruence (Tucker, 1951). This coefficient represents the cosine of the angle between two vectors, and in terms of its use in P-technique factor analysis, it measures whether the loadings are the same across individuals. This coefficient is described in the equation 2.1:

$$Q_{x_1,y_2} = \frac{\sum x_{i1}y_{i2}}{\sqrt{\sum x_{i1}^2 \sum y_{i2}^2}} \quad (2.1)$$



where:

$x_{i1}$  : Loadings for individual 1

$y_{i2}$  : Loadings for individual 2

$i : 1, \dots, p$

$p$  : Number of variables

A recommended cut-off value for this coefficient is 0.85 (Haven & ten Berge, 1977). So if the coefficient is higher than 0.85 then the loadings can be assumed to be the same across individuals, which implies that there is a similar change pattern between individual 1 and 2. Note that this coefficient is derived as a sum over all variables between a pair of individuals. So, this coefficient should be calculated between each individual and all the remaining participants of the study.

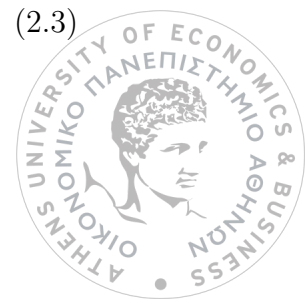
P-technique is a popular technique due to its simplicity and the notion that every individual process is best studied through intensive (high-volume) longitudinal measurements (P. C. Molenaar & Nesselroade, 2009). Such a method could be considered as an EFA model for time series data by performing the comparison within-individuals over time. The use of a time series model enables the investigation of lagged relationship between scores. For instance, it could be helpful for understanding how an individual level's stress today will affect the negative mood tomorrow, given that stress and negative mood are 2 domains of an instrument.

Although such a method is commonly used when there is a high volume of repeated measurements across individuals, in the area of PRO psychometric evaluation is not commonly used. This is because in the psychometric field, within-individual variability is mostly considered as statistical nuisance. This comes in contrast with the P-technique, which presumes that within-individual variability is an insightful source of variance that should be captured in a model.

This method has 2 main components: a factor analytic model to account for the measurement error and a time series model. The P-technique factor analysis model is described in more detail by equation 2.2 and 2.3:

$$y_t = \Lambda \eta_t + \epsilon_t \quad (2.2)$$

$$\eta_t = B \eta_{t-1} + \zeta_t \quad (2.3)$$



where:

$y_t$  :  $p$ -dimensional vector of observed time series at time  $t$

$\eta_t$  :  $k$ -dimensional vectors of the factor scores at time  $t$

$\eta_{t-1}$  :  $k$ -dimensional vectors of the factor scores at time  $t - 1$

$\Lambda$  :  $p \times k$  factor loading matrix which does not change over time

$B$  :  $k \times k$  matrix with the auto and cross regressive coefficients between and within factor scores at time  $t-1$  and time  $t$

$\epsilon_t$  :  $p$ -dimensional residuals for the factor analysis model

$\zeta_t$  :  $k$ -dimensional residuals for the time series model

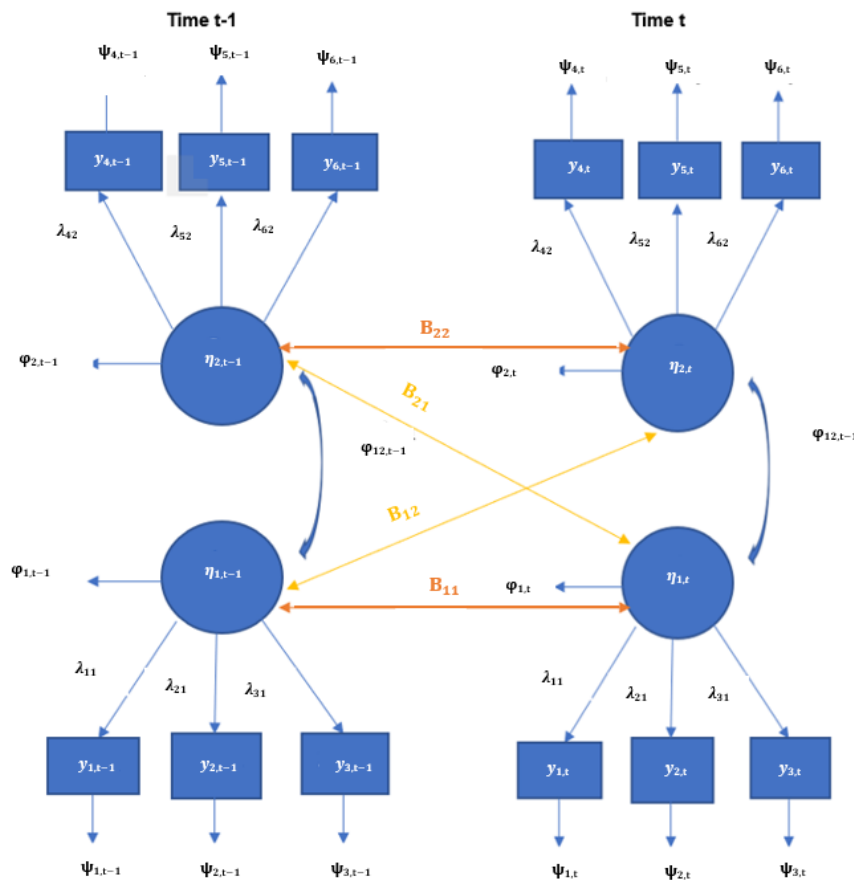
$p$  : Number of variables

$k$  : Number of factors

The model is displayed analytically on Figure 2.1.



Figure 2.1: Example of P-technique factor analysis model with 2 factors and 6 items.



$\lambda_{ij}$ : Loading for the  $i$  variable and  $j$  factor

$y_{it}$ :  $i$  variable at time  $t$

$\psi_{it}$ : Error variance for the  $i$  variable at time  $t$

$\eta_{jt}$ : Factor score for the  $j$  factor at time  $t$

$\phi_j$ : Factor variance of  $j$  factor

$B_{bc}$ : Correlation of the factor score  $b$  at time  $t-1$  and factor score  $c$  at time  $t$

Although this method is simple, it entails some limitations. Firstly, it does not account for the time-dependent nature of the factors and as a result the latent variable at time  $t$  affects the observed variables only at time  $t$  (I. A. Lee & Little, 2012). This means that there is not any lagged relationship between observed and latent scores. This is an important limitation when studying individual processes. Secondly, this method is not appropriate when the scalar does not hold (i.e., the number of factor, the items that are associated within each factor and the association strength among factors and items remain constant across time). Such method assumes the number of factors and the factor loadings are constant across time. This means that the factors exert the same amount of influence on the observed items across the study period, but this is not always true. That is why more advanced models which are called dynamic

factor models (P. Molenaar, 1985) have been proposed to overcome the limitation of P-technique models. Such models have been used for daily diary PRO data (H. Song & Zhang, 2014) but not very extensively. These models will not be under the scope of this thesis, so for more details see (I. A. Lee & Little, 2012; H. Song & Zhang, 2014; P. Molenaar, 1985; Hamaker & Wichers, 2017).

## 2.2.2 Design-based models

When the goal is not to study individual change but to employ a model that will take into account the intra-individual nested structure by using an overall model, a design-based approach is an appropriate method. This approach employs an overall model by assuming metric invariance across the levels (Kaplan & Elliott, 1997). In the case of daily diary data, this means that the number of factors and the loading on the between-individual level are assumed to be the same with the number of factors and loadings on the within-individual level. This approach was initially proposed to take into account the increased bias in the standard error of the fixed estimations of the model due to the dependency of the nested structure of the data (Neyman, 1992) by using robust standard errors (Huber, 1967) such as the well known sandwich-type variance estimator. This adjustment affects standard errors and not parameter estimates (Hardin, Hardin, Hilbe, & Hilbe, 2007) and they enable to make valid inferences. More specifically, this method comprises the estimation of an overall model with the ultimate goal of making more accurate inferences about the lower level of the data based on the adjusted standard errors. The estimation of the standard errors will either be estimated through linearization (Bentler, 2010), generalized estimated equations (Liang & Zeger, 1986) or replication methods such as jackknife and bootstrap (Efron & LePage, 1992).

This model is described in equation 2.4

$$y_i = \Lambda\eta_i + \epsilon_i \quad (2.4)$$

where:

$y_i$  :  $p$ -dimensional vector for the  $i$  observation

$\eta_i$  :  $k$ -dimensional vectors of the factor scores for  $i$  observation

$\Lambda$  :  $pxk$  factor loading matrix

$\epsilon_i$  :  $p$ -dimensional vector for the  $i$  residual

$p$  : Number of variables

$k$  : Number of factors



Note that this approach uses all the data-set and  $i$  index does not refer to the individual but to the index of the observation.

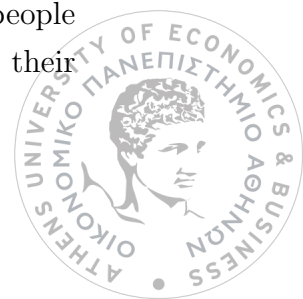
This method can yield satisfying results when the latent structure is assumed to be the same for the between- and within-individual level (Wu & Kwok, 2012). However, such an assumption may not always hold, making this model less appropriate in such instances. If a researcher is interested in answering questions with reference to either the between- or within-individual level, model based methods such as multilevel models are more appropriate, where there is construction of separate models of each level. Even though this method could be used in daily diary studies, it is not commonly used in the area of daily diary PRO measurements. It is often used to model complex survey data where data are collected via cluster or multistage sampling (e.g., some individuals are within the same class or household).

### 2.2.3 Independent analysis

Although taking into account the dependency of the data seems an appropriate approach, there is an alternative method that does not account for this significant component. This simplistic approach utilize all the data and ignores the intra-individual nested structure of the data by performing an independent analysis (Frison & Pocock, 1992). Such a method is not an efficient method to handle daily diary data, but it could be employed before conducting more complex approaches such as multilevel models that will be introduced. That is why this method is also described in this Chapter.

This technique is often called R technique (Cattell, 1963) under the framework of latent variable modelling. This approach is inappropriate as it leads to a wrong analysis as the data within each individual cannot be assumed to be independent due to their longitudinal nature (Bolger et al., 2003) but it can be useful as an initial step to get a first glance to the latent structure of the data (Reise et al., 2005). Treating individuals as if they were independent on a lower level could possibly lead to biased estimates and to an important violation when conducting a single-level analysis (J. Hox, 1998). These occur as the assumptions of normally distributed, independent and homoscedastic errors are violated (J. Little, 2013) as the intra-individual nested structure of the data leads to more complex errors on the within individual data.

Due to the nature of the data, the total variance of each item and the covariance between items is affected by the variation of item ratings within individual across time and by the variation between individuals based on their average differences. For instance, some people are in better mood on average in comparison with other people (between-individual variance), and some people are in better mood relative to their own typical level of mood (within-individual variation) (J. Hox, 1998).



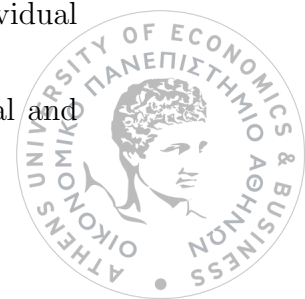
The consequences of ignoring the intra-individual nested structure was illustrated in a Monte Carlo simulation (Julian, 2001), where it was concluded that if intraclass correlation coefficients (ICC) (Koch, 2004) are greater than 0.05 the dependency of the data should not be ignored as there will be an increased bias in estimated parameters, their corresponding standard errors and chi-square statistic inflation. ICC refers to the amount of total item variance that is explained due to between-individual variance. Values close to 1 indicate that a high percentage of the total variance of the item is explained by the between source of variance, whereas values close to zero indicate that the total variation is explained mostly by the within source of variance, so values close to 1 indicate that the data are independent. Although independent analysis is not an appropriate method, it is an initial step before conducting a multilevel factor analysis (B. O. Muthén, 1991) approach as it allows the measurement of the dependency of the data for each item via ICC as the basic requirement for multilevel factor analysis is the presence of both between- and within-individual variation (Snijders & Bosker, 2011).

#### 2.2.4 Multilevel models

Multilevel factor analysis models could be considered as a typical method when it comes to daily diary studies (Bolger et al., 2003) and they enable the investigation of the data in multiple levels: within- and between-individual level. However, in the area of PROs they have not been extensively used, as more simplistic approaches are usually employed. The reason for its limited use might be attributed to its complexity and that within-individual variability is considered noise rather than an insightful source of variance. Such models try to account for the multilevel nature of the data as in the case of the design-based approach, but they capture the dependency through the construction of different models of each level of the data. These models estimate both within- (level 1) and between- (level 2) individual models.

Multilevel models are able to capture useful insights that can not be captured by simplistic approaches. For instance, an often-untested assumption is whether the psychometric properties of an instrument can generalize from the between-individual context to within-individual evaluations (Mehta & Neale, 2005; P. C. Molenaar, 2004). Simply assuming that the measurement properties, including the latent structure, are the same at both the within- and between-individual level in the absence of empirical evidence could result in ecological fallacy (Robinson, 2009). In some cases the sensitivity of items due to the within-individual differences might be larger than between-individual differences and this might imply different interpretation across the 2 constructs.

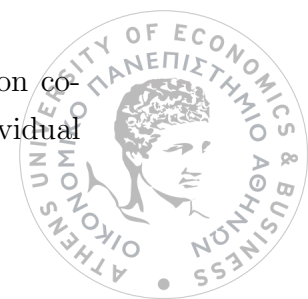
Multilevel models allow the checking of key assumptions such as dimensional and



cross-level invariance. The first property requires the number of factors to be the same at both the between- and within-individual level. This is an important property because if it does not hold, the items formulate different constructs across the two measurement levels (Gregorich, 2006) and may impact sensitivity of the measure differently at the two levels. The second property refers to the equality of the estimated factor loadings across both levels (Skrondal & Rabe-Hesketh, 2004). This is also an important property which ensures that the interpretation across the two levels is the same. Studies have shown that dimensionality at the within- and between-individual level can indeed differ (Roesch et al., 2010). Even when the assumption of dimensional invariance holds, cross-level invariance may not hold, as the item strength and sensitivity might not be the same when examining between-versus within-individual differences (Snijders & Bosker, 2011). Additionally, in the case where the number of individuals in a study is small, multilevel models could be quite useful as they utilize a greater proportion of data in comparison with other approaches that utilize smaller sample size (Houts, Morlock, Blum, Edwards, & Wirth, 2018).

In the evaluation of daily diary data, both between- and within-individual sources of variance can provide useful insights into specific components of the total variance. For instance, assessing the within-individual source of variance might aid in understanding the daily variability in the condition. As an example of this, Zautra et al., when studying fatigue amongst women with one of three forms of chronic illness, found that the three patient groups had differences in both their average level of fatigue and day-to-day variability, which exemplifies that useful insight can be retrieved by taking into account both between- and within-individual variability in such comparative research (Zautra et al., 2007). Yet, within-individual variance is often not considered in factor analysis. When using the weekly item average or the single selected day in cases where daily variability has a limited impact, the within-individual variance would not be insightful and will be considered noise. In this case, the weekly item average method and single selected day method will theoretically provide the same insights as would be attained when considering both the within- and between-individual variance separately (Schoemann, Rhemtulla, & Little, 2014). However, in cases where both sources of variance explain a significant proportion of the total data variation, the weekly item average method and single time point will result to loss of information. Given that it is not always possible to know a priori whether the within-individual variability is insightful or not, ignoring it seems not appropriate. Therefore, it is important to acknowledge the need of a measurement model such as multilevel model that takes into account the multilevel nature of the data.

The general goal of multilevel modelling is to decompose the total population covariance matrix into two components: a within-individual and a between-individual



covariance matrix (B. O. Muthén, 1994). The first level of the multilevel model (corresponding to the within-individual variability across days) reflects the relationship among item responses for the same individual over time, pooled across individuals. So, if for example two items are highly correlated in the within-individual matrix, this indicates that an increase/decrease of the value of the first item (relative to an individual's mean on this first item) will also occur on the second item (relative to an individual's mean on this second item). The second level of the model reflects the relationship among the items on average when considered over the study period as a whole. If two items are highly correlated in the between-individual covariance matrix, this indicates that two individuals who report high values on average on one item also tend to report high values on average on the other item. Both the within- and between-individual levels are analysed in the multilevel framework under the assumption that both provide useful and important insights.

Sample pooled within-individual covariance matrix:

$$S_{pooled} = \frac{\sum_{g=1}^G \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_{ig})(y_{ig} - \bar{y}_{ig})'}{N - G} \quad (2.5)$$

where:

$G$  : Total the number of individuals

$N$  : Total number of observations

$n_g$  : Number of observations within the  $g$  individual

$y_{ig}$  :  $i$  observation for the  $g$  individual

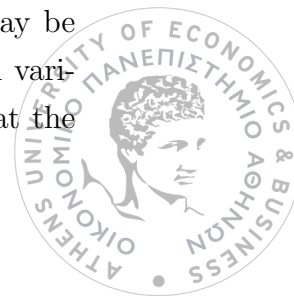
$\bar{y}_{ig}$  : Average value across time for individual  $g$

Sample between-individual covariance matrix:

$$S_{between} = \frac{\sum_{g=1}^G (\bar{y} - \bar{y}_{ig})(\bar{y} - \bar{y}_{ig})'}{G - 1} \quad (2.6)$$

where  $\bar{y}$  the total mean across all individuals and  $\bar{y}_g$  is the individual's mean across time.

Multilevel models in general have more error terms in comparison with single level models, which leads to more flexibility in defining the covariance structure. This increased flexibility allows researchers to investigate a variety of research questions regarding different parts of the covariance structure. For instance, researchers may be interested in examining how sensitive are the items due to the within-individual variation. Further, the hypothesis that there is equivalence in the latent structure at the

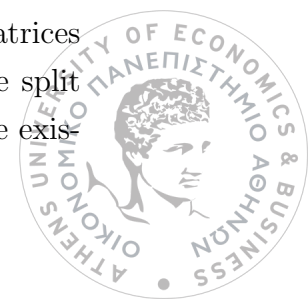


between- and within-individual level can be tested. This is important as it clarifies whether similar interpretation across both levels (i.e., at the group and individual-level) is appropriate. Although this kind of flexibility could be insightful, it can also pose great challenges within a clinical trial where one should prespecify the scoring algorithm of an instrument both at the within-person and between-person a priori.

An initial model that was firstly proposed to handle data with 2 or more levels was the Härnqvist model (Härnqvist, 1978), which decomposes the total score into 2 orthogonal scores: the between and within based on Cronbach and Webb's decomposition (Cronbach & Webb, 1975). They proposed to decompose the individual data  $Y_{ig}$  into the between-individual component  $Y_B = \bar{Y}_g$  and the within-individual component  $Y_W = Y_{ig} - \bar{Y}_g$ . So the total score  $Y_{ig}$  is replaced by  $Y_W$  and  $Y_B$ . This decomposition could also be employed in order to compute the between-individual covariance matrix  $\Sigma_B$  (covariance matrix of an individual's mean across a study period) and within-individual covariance matrix (covariance matrix of an individual's deviation relative to his own mean across a study period). This means that  $\Sigma$  which is equal to the total covariance matrix of the data is equal to  $\Sigma_B + \Sigma_W$ . This logic also applies for the sample data, where the sample total covariance matrix will be decomposed into the sample between-individual covariance matrix and within-individual covariance matrix. Although this was a great initiation for the multilevel models, it was only used for explanatory purposes, as it doesn't allow for statistical inference.

This notion regarding the decomposition of the total correlation matrix was also described in Muthén's (1991) approach. The goal of this approach is to decompose the variation of the observed variables into variance components associated with each level of the intra individual nested data and explain the variation of each level (B. O. Muthén, 1994). This enables the evaluation of both within- and between-individual levels and allows the evaluation of invariance in both levels (McNeish, Mackinnon, Marsch, & Poldrack, 2021).

This multi-stage approach can be used to evaluate latent structural properties of a multilevel assessment, such as in the case of daily diary measures. The procedure is to (i) firstly conduct the common factor analysis on the total correlation matrix by ignoring the dependency of the data, then (ii) to calculate the ICC to assess whether multilevel factor analysis is appropriate or not, and finally (iii) to conduct factor analysis for the within- and between-individual correlation matrices separately. Analysing the two correlation matrices separately could be beneficial as the two single level models can be assessed for adequacy (Yuan & Bentler, 2007), whereas evaluating model fit for the entire model (i.e., which can be achieved by estimating both covariance matrices simultaneously) may cause several problems (Lindqvist et al., 2017). When the split modelling approach is deemed adequate, researchers can draw guidance from the exist-



tent literature on standard structural equation modelling (Bentler, 2010) for analysing the between-individual and within-individual variance.

However, where such adequacy is not established, and a multilevel approach is utilized where both within- and between- individual parameters are estimated simultaneously, then there is no information on whether the poor fit is due to the first- or second-level model. Additionally, the fit indices are less sensitive to misfit in the between-individual model than the within-individual. This is because in most cases the number of observations at the within-individual level is much larger than the between level, so the within-individual model has more weight. When that is not the case, the between-individual model has higher weight than the within-individual model in the level of the misfit of the model. Another issue is that if there is a specification on one level, then the parameter estimates of the other level will be likely influenced. Consequently, when there is specification in the model at either level, even if it is small, the results may indicate poor model fit overall (Yuan & Bentler, 2007).

In terms of implementation, the sample (pooled) within-individual covariance matrix is considered an unbiased estimate for the population within-individual covariance matrix and do not cause any modelling challenges (Huang, 2017). It has been empirically shown that factor analysing only the within-individual variability yields the same results with the within-individual parameters as would be produced by a multi-level model that models the parameters for both levels simultaneously (B. O. Muthén, 1994).

Although there are several advantages when using the split approach, there is an important limitation. The estimated between-individual covariance matrix is not an unbiased estimator of the population between-individual covariance matrix (B. O. Muthén, 1991). Instead, this covariance matrix is an unbiased estimator of a linear combination of the population within- and between-individual matrices (J. Hox & Maas, 2004). Although an alternative unbiased estimator of population between-individual matrix has been produced (Heck, 2001), it is not employed as it is usually non-positive finite (B. O. Muthén, 1994).

The sample between-individual covariance matrix  $S_b$  is an unbiased estimator of the sum of the population within-individual covariance matrix  $\Sigma_W$  and the population between-individual covariance matrix  $\Sigma_B$  multiplied by the term  $c$  which represents the average cluster size (B. O. Muthén, 1994). Therefore, the sample between-individual covariance matrix comprises 1 unit of within-individual variance and  $c$  units of between-individual variance. For balanced designs, the value of  $c$  is equal to the number of observations within each individual and, for unbalanced designs,  $c$  is computed as described in equation 2.7.



$$c = \frac{N^2 - \sum_{g=1}^G n_g^2}{N(G-1)} \quad (2.7)$$

where:

$N$  : Total number of observations

$G$  : Total number of individuals

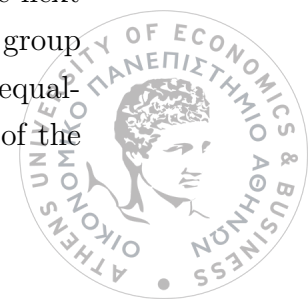
$n_g$  : Number of observations within the  $g$  individual

$y_{ig}$  :  $i$  observation for the  $g$  individual

The benefits of factor analysing both within- and between-individual is that it can address the chi-square inflation that arises due to non-independence. However, it is mainly based on normal theory estimators (e.g., maximum likelihood), so it is unable to address chi-square inflation, parameter bias and standard error attenuation in case of ordered data (Finney & DiStefano, 2013).

Given the computational challenges of the between-individual model which requires the prior specification of the within-individual model, multilevel models have been developed which estimate the within- and between-individual covariances matrices simultaneously. For the between part of the model, this approach necessitates the inclusion of both within- and between-individual covariance matrices in a multigroup set-up based on  $n-G$  and  $G$  observations respectively, where (as specified in equation 2.7)  $n$  denotes the total number of observations and  $G$  denotes the total number of individuals. This means that the model for the population-level within-individual covariance matrix is defined in both levels of the model. As a consequence, equality constraints should be imposed to ensure that the model for  $\Sigma_W$  is the same across the two levels. On the other hand, the model for  $\Sigma_B$  is specified for the  $S_b$  only, along with the scaling factor  $c$  which is used in the model development. In the within-individual part of the model, things are simpler, as only the sample pooled covariance matrix  $S_{pw}$  is used.

The notion that the within-individual covariance matrix is required for the construction of both levels of the models was conceptualized in the Hox approach (J. J. Hox, Moerbeek, & Van de Schoot, 2017). The step 1 for this procedure is to conduct a single level factor analysis only for the  $S_{pw}$  and ignore the  $S_b$ . This allows the evaluation of whether the fit of the within-individual model is adequate and whether no further exploration needed for the factor structure of the between-individual model. The next step is to calculate the NULL model where both  $S_{pw}$  and  $S_b$  are used in a multi group set up by using the factor structure on both matrixes as defined in level 1 with equality constraints. Equality constraints are essential, as  $S_{pw}$  appear in both levels of the



model. These constraints include equality of factor loadings, equality of variances and covariances of the observed variables and the factors across both levels. The third step, which necessitates the estimation of the independence model, requires to calculate the between portion of variance of the model by adding the between-level factors (but not covariances). This requires the use of the average cluster size  $c$  (or scaling factor) as each observed variable comprises one unit of the  $\Sigma_w$  and  $c$  units of  $\Sigma_b$  variance. This means that the between-level variables are not fixed to 1 as it is usually assumed but equal to  $\sqrt{c}$ . This is important as this scaling factor transforms the group level variables in the proper scale. The fourth step requires the estimation of a saturated model where there is no restriction for the between-part of the model in order to assess its fit. It is expected that if the null model had an adequate fit, then the saturated should also have a good fit too. The final step requires the estimation of the hypothesized model, which require the factor structure of both within- and between-individual model and the covariances of the group variables, in contrast with independent model where only the variances of the group level variables is available. Assuming that we have a 2-factor within- and a 2-factor between-individual model with 4 items ( $x_1, x_2, x_3, x_4$ ), the single level, independence and the hypothesized model are displayed by the below path diagrams:

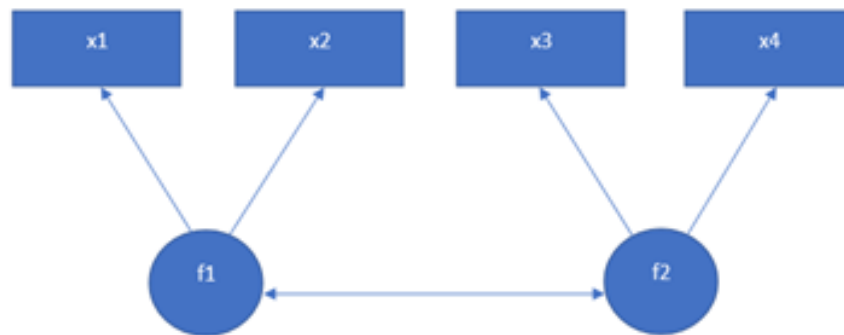


Figure 2.2: Single level within-individual model with 2 factors and 4 items

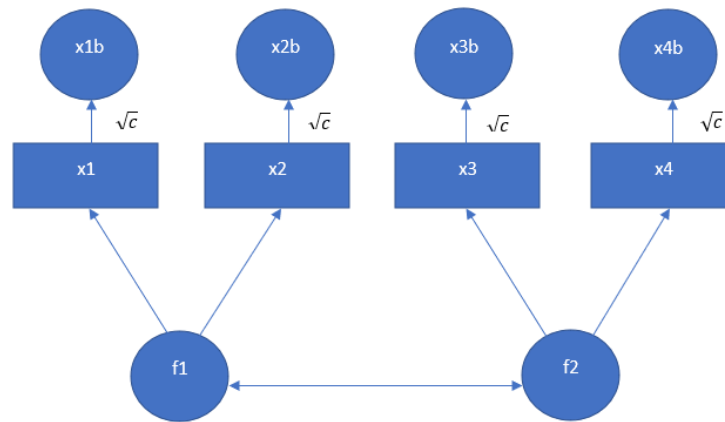


Figure 2.3: Independence model

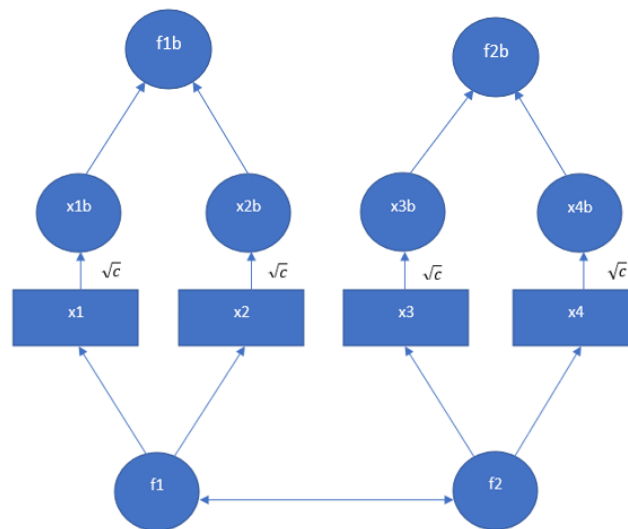


Figure 2.4: Hypothesized model

These models as previously flagged are constructed based on the notion that that total covariance matrix is decomposed into between and within-individual matrix. This also applies for the factor and residual covariance matrix.

Let  $y_i$  be a  $p$ -variate vector  $y_i$  for individual  $i$  and  $y_i$  is multivariate normal distributed.

Then:

$$y_{g_i} = v + \lambda\eta_{g_i} + \epsilon_{g_i} \quad (2.8)$$

where:

$v$  : Intercept vector

$\lambda$  : Vector for factor loadings

$\eta$  : Factor score

$\epsilon$  : Residual vector

$g$  : Subscript for individual  $g$

$i$  : Subscript for the observation nested in individual  $g$

$$\eta_{gi} = a + \eta_{B_g} + \eta_{W_{gi}} \quad (2.9)$$

where:

$a$  : Overall mean for  $\eta_{gi}$

$\eta_{B_g}$  : Random factor component to capture individual effect

$\eta_{W_{gi}}$  : Random factor component varying for the same individual across time

The decomposition of total factor variance will be :

$$V(\eta_{gi}) = \psi_T = \Psi_B + \Psi_W \quad (2.10)$$

where:

$\eta_{gi}$  : Factor score for the  $g$  individual and  $i$  observation

$\Psi_B$  : Variance of the between-level factor score

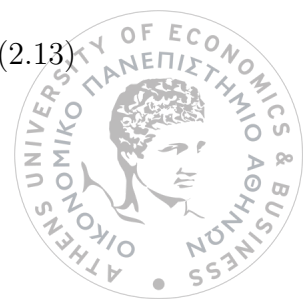
$\Psi_W$  : Variance of the within-level factor score

The residual variance will also be decomposed into two part:

$$V(y_i) = \Sigma_T = \Sigma_W + \Sigma_B \quad (2.11)$$

$$\Sigma_B = \lambda_B \Psi_B \lambda_B' + \Theta_B \quad (2.12)$$

$$\Sigma_W = \lambda_W \Psi_W \lambda_W' + \Theta_W \quad (2.13)$$



here:

$\Sigma_{S_T}$  : Total covariance matrix

$\Sigma_W$  : Within-individual covariance matrix

$\Sigma_B$  : Between-individual covariance matrix

$\lambda_W$  : Within-level loadings matrix

$\lambda_B$  : Between-level loadings matrix

$\Psi_W$  : Within-level error variance matrix

$\Psi_B$  : Between-level error variance matrix

$\Theta_W$  : Within-level error variance matrix

$\Theta_B$  : Between-level error variance matrix

Although the multilevel enables the investigation of between- and within-individual covariance matrix, in some applications the total variation is of primary interest. Even in this case, multilevel models can provide the estimate of total covariance matrix as the sum of the estimated between- and within-individual covariance matrix and make valid inference. However, there are more simplistic approaches that could make adjustment for the standards errors and chi-square statistics (B. Muthén & Satorra, 1989) as was showed in the case of designed based approach.

To sum up, the multilevel models can be described by the below equations:

$$y_{g_i} = v + \lambda_B \eta_{B_g} + \lambda_W \eta_{W_{g_i}} + \epsilon_{B_g} + \epsilon_{W_{g_i}} \quad (2.14)$$

$$V(y_i) = \Sigma_T = \Sigma_W + \Sigma_B \quad (2.15)$$

$$\Sigma_B = \lambda_B \Psi_B \lambda_B' + \Theta_B \quad (2.16)$$

$$\Sigma_W = \lambda_W \Psi_W \lambda_W' + \Theta_W \quad (2.17)$$

For more general forms of multilevel models, see (Schmidt & Wisenbaker, 1986; B. O. Muthén, 1989; McDonald & Goldstein, 1989).



# Chapter 3

## Simulation study

### 3.1 Introduction

Multilevel models serve as a useful methodology to analyze daily diary PRO data. Consequently, such data entail a variety of data handling and modelling options under factor analysis framework. That is the reason for which it is important for someone to consider the proper method based on the goal of the study. In daily PRO data, particularly single-day and item average approach have been used to explore the properties of the latent structure of a daily diary instruments and their appeal is attributed to their simplicity and the fact that within-individual variability is considered as a statistical nuisance. However, that is not always the case as multiple researchers have supported that within-individual variability could in fact provide useful insights when it comes to different type of psychological processes as it is shown in Table 3.1



Table 3.1: Summary of the literature regarding the within-individual variability in psychological processes and symptoms.

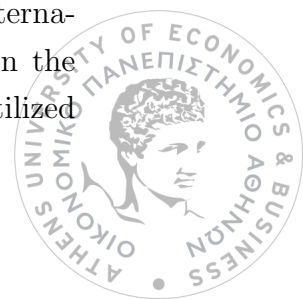
<b>Within-individual variability in psychological processes and symptoms.</b>	<b>Researchers</b>
Importance of within-individual variability when studying personality.	(Hooker, 1991; Shoda, Mischel, & Wright, 1994)
Importance of within-individual variability when studying affect, emotion mood.	(Larsen, 1987; Lebo & Nesselroade, 1978; Wessman & Ricks, 1966; Zevon & Tellegen, 1982)
“Within person variability has been found to relate across domains of cognition and psychological function in persons with dementia.”	(Hoffman, 2007)
“Between-individual differences in variability is an important but neglected topic in psychological research.”	(Greenier et al., 1999)
“Diary studies have showed that there is moment to moment and day to day fluctuations in symptoms.”	(Cranford et al., 2006; Schneider et al., 2012)
“Within-individual variability is often considered a statistical nuisance, even though it could be valuable, as it could be considered an important construct itself.”	(Nesselroade & Ram, 2004; Ram & Gerstorf, 2009)

Multilevel models on the other hand do not consider the within-individual variability as a nuisance but as an insightful source of variation with its own construct. So when within-individual variability is a valuable source of information, the simplistic approaches lead to loss of information.

For this reason, in this thesis, a simulation study was conducted to investigate whether item average and single day approach are prone to methodological issues under EFA and CFA framework. An additional goal was to elucidate the need of multilevel factor analysis model when both within- and between-individual variances are of interest and address its possible limitations.

When it comes to diary PRO instruments, items are provided on a categorical format, rather than continuous. For this reason, an IRT model was utilized for simulating ordered data in likert scale (0-4). Multidimensional graded response model in particular was the model under which this simulation study was conducted.

For the single selected day approach where the data were on an ordinal scale, ordinal EFA and CFA were used, whereas for the item average approach common EFA and CFA were utilized. EFA and CFA multilevel models were also used as alternative ways to explore the latent structure of the daily diary instrument. Within the exploratory framework as applied in the present study, the split approach was utilized



to allow the application and comparison of common criteria for selecting the number of factors across each of the data handling strategies under investigation. Within the confirmatory framework, where the number of factors is prespecified, a multilevel approach was utilized in which the within- and between-individual models were estimated simultaneously.

In this Chapter the simulated model, the equivalence of the IRT with CFA model, the modelling methods for each approach and the assessment methods to evaluate the methodological issues under both EFA and CFA will be presented.

## 3.2 Description of the simulation study

A simulation mechanism was implemented under a three-factor multidimensional graded response model due to its flexibility (Depaoli, Tiemensma, & Felt, 2018) as it allows some of its parameters to vary across the items and item levels. The simulated datasets were designed to represent the results from a questionnaire comprising 20 items as if it was administered daily over a 1-week period. The 20 items were scaled from 0 to 4, the day-to-day correlation among factor scores was set to 0.85. The correlation between items, slope parameters and threshold parameters were set to be constant across time. The first 10 items labelled as “item 1”, “item 2”, . . . , “item 10” load on the same factor which is labelled as “factor 1”, and the next 5 items labelled as “item 11”, “item 12”, . . . , “item 15” load on the same factor which is labelled as “factor 2” and the last 5 items labeled as “item 16”, “item 17”, . . . “item 20” load on the same factor which is labelled as “factor 3”.

The day-to-day correlation among factors scores is described in the equation 3.1. For  $t=1$ :

$$\theta_1 \sim N \left[ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix} \right]$$

where:

$\theta_1$  : 3x1 vector of the latent scores for day 1



For  $t=2,3,4,5,6,7$ :

$$\theta_t = 0.85\theta_{t-1} + \sqrt{1 - 0.85^2}x \quad (3.1)$$

where:

$\theta_t$  : 3x1 vector of the latent score at time  $t$

$t$  : Index for the number of day

$x$  : 3x1 vector of a random variable which follows the same distribution with the latent score at day 1

The multidimensional graded response model each day ( $t=1,2,3,4,5,6,7$ ) is described by equation 3.2, 3.3 and 3.4.

$$P(x_{tij} \geq c|\theta_{tik}) = \frac{1}{1 + D \exp \sum_{k=1}^3 a_{jk}(\theta_{tik} - d_{jc})} \quad (3.2)$$

$$P(x_{tij} \geq 0|\theta_{tik}) = 1 \quad (3.3)$$

$$P(x \geq 5|\theta_{tik}) = 0 \quad (3.4)$$

where:



$t$  : Day that the patient answered the questionnaire

$i$  : Subscript for the individual ( $i=1, \dots, G$ )

$j$  : Subscript for the items ( $j=1, \dots, 20$ )

$c$  : Item levels of the items that range from 0-4 (Likert-scale)

$k$  : Subscript for the factors ( $k=1, 2, 3$ )

$x_{tij}$  : Response to the questionnaire at time  $t$  for the  $i$  patient on  $j$  item

$a_{jk}$  : Slope parameter for the  $j$  item and  $k$  factor which indicates how well an item can discriminate between subjects with differences in the latent trait

$d_{jc}$  : Difficulty parameter for item  $i$  and item level  $c$  which represents the point on the latent variable separating category  $c$  from category  $c+1$

$\theta_{tik}$  : Latent trait at time  $t$  for  $i$  individual and factor  $k$

$D$  : Scaling factor which is set to be equal to 1.702 so that the logistic metric approximates the normal ogive model (Reckase, 2009)

Equation 3.3 and 3.4 represent the probability of patient  $i$  to respond to item  $j$  at time  $t$  an item category greater than zero and 5 correspondingly. The total scenarios under which the simulation study will be conducted is 2, and they will be implemented under various sample sizes. Table 3.2 provides a detailed presentation of the different combinations of the parameters that will be employed.

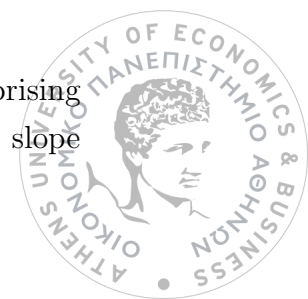
Table 3.2: Scenarios of the simulation study based on the values of slope parameter, and sample size.

Slope parameter	Sample size
Scenario 1: $1.35 \leq a \leq 1.69$	$n=100, 150, 200, 250, 350$
Scenario 2: $a \geq 1.7$	$n=100, 150, 200, 250, 350$

Based on the range of the slope parameters that were selected:

- When the  $a \geq 1.7$  the performance of the items is satisfactory (Baker, 2001)
- When  $1.35 \leq a \leq 1.69$  the items are functioning well and little or no revision is required (Baker, 2001)

The simulation mechanism was implemented to generate 1,000 datasets each comprising  $n=100, 150, 200, 250, 350$  observations with different options for the values of the slope



parameter ( $a$ ) as described in Table 3.2, where small sample sizes were selected as being reflective of typical Phase 2 clinical trials where PRO instruments are commonly utilized and assessed. The slope and threshold parameters for each day of observation were assumed to be constant, as was previously mentioned, such that scalar invariance held across time (McNeish et al., 2021; Putnick & Bornstein, 2016), an important property when it comes to the evaluation of the daily diary instrument. Finally, the data for this simulation study were simulated based on flexMIRT package in R (Chalmers, 2012),

### 3.3 Equivalence of graded response model with-Confirmatory factor analysis model

The analysis in the present study was mainly focused on using EFA and CFA, so the estimated loadings and estimated correlation matrix were one of our main targets of evaluation. For this reason, the known equivalence of IRT models with categorical CFA models (Takane & De Leeuw, 1987) was used to derive simulated loadings and correlation matrix.

Based on their equivalence, the loadings, and correlation matrix were specified as explained in equations 3.5 and 3.6.

$$L_{jk} = \frac{\frac{a_{jk}}{D}}{\sqrt{1 + \left(\frac{a_{jk}}{D}\right)^2}} \quad (3.5)$$

$$R = LCov(\eta)L' + \psi \quad (3.6)$$

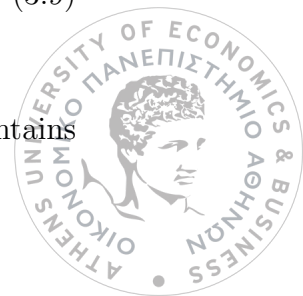
where:

$$Cov(\eta) = \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \quad (3.7)$$

$$L = \begin{bmatrix} L_{1,1} & L_{1,2} & L_{1,3} \\ L_{2,1} & L_{2,2} & L_{2,3} \\ \dots & \dots & \dots \\ L_{20,1} & L_{20,2} & L_{20,3} \end{bmatrix} \quad (3.8)$$

$$\Psi = \begin{bmatrix} 1 - L_{1,1}^2 & 0 & 0 & \dots & 0 \\ 0 & 1 - L_{2,1}^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 - L_{3,20}^2 \end{bmatrix} \quad (3.9)$$

$Cov(\eta)$  is a  $3 \times 3$  correlation matrix of the factors,  $L$  is a  $20 \times 3$  matrix which contains



the loadings of items within each of the 3 factors and  $\Psi$  is the uniqueness, and it is the part of variance that remains unexplained by the model, and it is calculated based on delta parametrization (Brown, 2015):

$\Psi$  is a diagonal matrix  $20 \times 20$  and for  $j=1, \dots, 20$ , each diagonal value contains the following quantity:  $1-L_j^2$ .

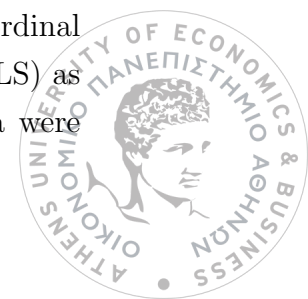
## 3.4 Methods for exploratory factor analysis

### 3.4.1 Modelling methods

EFA was conducted under three different approaches for the simulated data. In the first approach, a single timepoint (specifically Day 5 of 7) was selected and polychoric correlation coefficients were utilized as the input matrix for the EFA. In the second approach, average of the item score across the week was computed, and given the continuous scale of the derived values, Pearson correlations were used for the input correlation matrix. The third approach sought to account for both within- and between-individual sources of variance using the data-based Muthén approach (B. O. Muthén, 1994). A single-level factor analysis was implemented for each level in the multilevel approach, so existent literature and guidance regarding single-level methodologies was applicable.

This applicability was ensured as the third approach was implemented for the common EFA method using the sampled pooled within-individual covariance matrix (see equation 2.5) as input covariance matrix for the within-individual analysis and the sample between-individual covariance matrix (see equation 2.6) as the input covariance matrix for the between-individual analysis. Implementation of this third approach was, however, based on the presumption that both within- and between-individual sources of variance should be present in the data. This assumption was checked via ICC (Koch, 2004). ICC values close to 0 were considered indicative that almost all variability in the data was explained at the within-individual level, whereas ICC values close to 1 were considered indicative that almost all variability was explained by the between-individual variability. More specifically, inference was guided by recommended thresholds ( $<0.05$  and  $>0.95$ , respectively; (Schoemann et al., 2014)). After checking the presence of the multilevel variance in the data by computing the ICC for each observed variable across the time interval of interest, an EFA was then conducted for each of the ‘pooled within’ and ‘between’ individual covariance matrices.

For the single selected day approach where the data were on an ordinal scale, ordinal factor analysis was conducted by using Diagonally Weighted Least Square (DWLS) as an estimation method. For the weekly item average approach where the data were



continuous, maximum likelihood (ML) was used as an estimation method. As for the split EFA approach, both within-individual and between-individual covariance matrices were factor analysed by using ML as an estimation method. All EFA approaches were implemented via lavaan package in R (Rosseel, 2012).

### 3.4.2 Assessments methods

Estimates of interest in EFA included the number of factors identified, the loadings of the items on those factors, and correspondence between the estimated, observed and true correlation matrices, bias and mean square error (MSE) of the estimated loadings and goodness of fit measures.

For both single selected day and weekly item average the following assessment methods were applicable:

- The number of factors
- The magnitude and range of loadings
- Factor loadings bias and MSE
- The estimated, true and observed inter item correlation
- Goodness of fit measures

Number of factors was addressed across all iterations of the simulation study to identify whether there are any selection criteria that may be prone to high frequency of incorrect or correct identification for either the single selected day or weekly item average. The magnitude, range, bias and MSE of the loadings were addressed in order to examine whether ignoring within-individual variability could impact the results of the exploratory factor analytic model. The estimated true and observed inter item correlation were also assessed to examine to what extent the use of a proportion of the data (single selected day approach) or the aggregation of the data (weekly item average) differ in terms of their insights they provide for the inter item correlation. Goodness of fit measures were also utilized to examine the performance of the item weekly average and single selected day approaches under 4 different goodness-of-fit measures. For the split (multilevel) modelling approach, the following assessment methods were targets of evaluation:

- The number of factors
- The magnitude and range of loadings
- Goodness of measures



Number of factors was estimated for the same reasons as in the 2 previous approaches. The magnitude and range of loadings were investigated in order to examine whether the items show the same amount of sensitivity due to both within- and between-individual variability. Goodness of fit measures were also utilized, as in the case of weekly item average and single selected day. Bias, MSE and correlation of items were examined for the selected day and weekly item average approaches to obtain insight into the information loss that may occur when data were subject to data reduction/transformation that masked within-individual variability. Accordingly, these targets of assessment were not applicable in the case of multilevel models, where such variability is taken into account.

### **Number of factors**

As exploratory modelling presumes no a priori knowledge regarding the optimal number of factors in the measurement model, it is necessary to explore how many factors are present in the data. In the present study, a series of criteria were used in assessing the number of factors identified. The outputs of these criterion assessments were used to determine the percentage of analytic iterations for which the correct number of factors are identified, given that the target of this simulation study is three factors. The criteria used were: (i) the Kaiser criterion (Kaiser, 1960), (ii) parallel analysis (Horn, 1965) and (iii) the empirical Kaiser criterion (Braeken & Van Assen, 2017).

### **Factor loading strength and range**

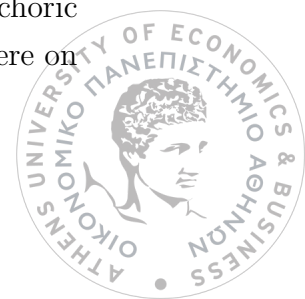
Another goal was to check whether the items loaded on their corresponding factor based on a specific threshold value of 0.45 (MacCallum, Widaman, Zhang, & Hong, 1999). This criterion was applied to all modelling approaches implemented in this study. Factor loading strength was also visualized through boxplots based on their distribution across 1,000 iterations of the simulation study. The range of the estimated loadings within each factor was also provided.

### **Estimated, true and observed inter item correlation**

The range for the observed, estimated inter item correlations was additionally explored, and it was compared with the true correlation matrix under the hypothesized model.

### **Observed correlation matrix**

In the single selected day approach where the items were in ordinal scale, polychoric correlation was utilized. In the weekly item average approach where the items were on a continuous scale, Pearson correlation was used.



### Estimated correlation matrix

In the single selected day approach, the estimated correlation matrix was estimated via the help of ordinal factor analysis model. In the weekly item average approach, it was calculated based on the common factor analysis model.

### True correlation matrix

The true correlation matrix was calculated based on the equivalence of IRT and CFA models as described in Section 3.3.

### Overall absolute factor loading bias and mean square error

Overall absolute bias of the estimated loadings (as defined in equation 3.10) was examined, as was the overall mean squared error (MSE) of estimated loadings (as defined in equation 3.11) for the weekly item average and the single selected day approaches.

$$Bias(\hat{\theta}) = \frac{\sum_{j=1}^{20} \frac{\sum_{N=1}^{N'} |\hat{\theta}_{jN} - \theta_j|}{N'}}{20} \quad (3.10)$$

$\hat{\theta}_{jN}$  : Estimated loading for the item  $j$  and the  $N^{th}$  replication of the simulated data

$\theta_j$  : True value of the loadings based on the equivalence of the IRT model and CFA model  
(see equation 3.5)

$N'$  : Number of times the model converged across 1,000 replications

$$MSE(\hat{\theta}) = \frac{\sum_{j=1}^{20} \frac{\sum_{N=1}^{N'} (\hat{\theta}_{jN} - \theta_j)^2}{N'}}{20} \quad (3.11)$$

where  $\hat{\theta}_{jN}$  and,  $\theta_j$  are defined as in the case of the bias of the estimated loadings.

### Goodness of fit measures

Finally, goodness of fit measures were estimated for each implementation of the three data handling approaches explored in this study under an exploratory modelling framework: Tucker Lewis Index (TLI), root mean square error of approximation (RMSEA), standardized root mean square residual (SRMR) and comparative fit index (CFI). The recommended cut-off values for each are provided in Table 3.3. For more details, see (Schermelleh-Engel, Moosbrugger, Müller, et al., 2003).

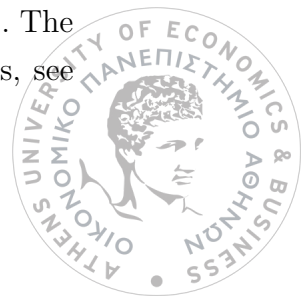


Table 3.3: Cut off values for CFI, TLI, RMSEA and SRMR.

Fit index	Cutoff values	
	Good fit	Acceptable fit
CFI	$0.97 \leq CFI \leq 1$	$0.95 \leq CFI < 0.97$
TLI	$0.97 \leq TLI \leq 1$	$0.95 \leq TLI < 0.97$
RMSEA	$0 \leq RMSEA \leq 0.05$	$0.05 < RMSEA \leq 0.08$
SRMR	$0 \leq SRMR \leq 0.05$	$0.05 < SRMR \leq 0.1$

### Convergence issues and Heywood cases

Percentage of convergence and occurrence of Heywood case were also reported in the results for all the approaches.

## 3.5 Methods for confirmatory factor analysis

### 3.5.1 Modelling methods

After conducting EFA, the next step was to conduct CFA with these 3 data-handling approaches: (i) selecting a single day of observation, (ii) utilizing weekly item averages, and (iii) conducting multilevel CFA by estimating both within- and between-level parameters simultaneously. The objective for the CFA was to estimate item loadings, item correlation matrices and goodness of fit measures. Models were estimated by setting factor variance equal to 1 for the single selected day and weekly item average approaches, while for the multilevel CFA the first item loading within each factor was set equal to 1 in order to avoid identification issues. Percentage of convergence and Heywood case was also reported as in the case of EFA.

For the single selected day approach where the data were on an ordinal scale, ordinal factor analysis was conducted by using DWLS as an estimation method whereas for the weekly item average approach where the data were continuous, ML was used as an estimation method. As for the multilevel CFA, ML was employed. All CFA approaches were implemented via lavaan package in R (Rosseel, 2012).

### 3.5.2 Assessments methods

Estimates of interest in the CFA stage of this work included the loadings of the items on their factors, correspondence between the estimated, observed and true correlation matrices, bias and MSE of the estimated loadings, and goodness of fit measures.

For both single selected day and weekly item average the following assessment methods were examined:



- The magnitude and range of loadings
- The estimated, true and observed inter item correlation
- Factor loadings bias and MSE
- Goodness of fit measures

For the multilevel modelling approach, the following assessment methods were applicable targets of evaluation:

- The magnitude and range of loadings
- Goodness of fit measures

All the above assessment method were addressed for the same reasons as addresses in EFA.

### **Factor loading strength and range**

In the CFA stage of this work, factor loading strength across all approaches were visualized through boxplots based on the distribution of the estimated loadings across 1,000 iterations of the simulation study. Additionally, the range of the estimated loadings within each factor was also reported for all approaches.

### **Estimated, true and observed inter-item correlation**

The same approach as used in evaluating the EFA modelling was implemented in examining the inter-item correlations (see Section 3.4.2).

### **Overall absolute factor loading strength and range**

The same formulas were used as in the EFA implementations for calculating the overall absolute average bias and overall MSE of the loadings (see Section 3.4.2).

### **Goodness of fit measures**

Evaluation of the fit of all the approaches was implemented. The same fit indices used in EFA were used as well in CFA (see Section 3.4.2).

### **Convergence issues and Heywood cases**

Percentage of convergence and occurrence of Heywood cases were also reported, as in the case of EFA.



# Chapter 4

## Exploratory factor analysis

### 4.1 Introduction

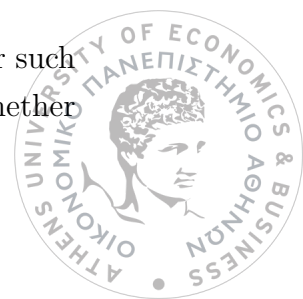
EFA is a multivariate statistical method that can be used when there is not a priori knowledge regarding the structure of a measure or the number of factors it may contain. EFA can assess the pattern of inter-item relationships and help identify the factors to which items may belong. In other words, it is a useful tool for evaluating the dimensionality of a multi item questionnaire. It thus can play an important role in questionnaire development, including daily diary PRO instruments. The purpose is to determine the appropriate number of factors that can explain adequately the inter item correlation. So if someone is interested to explore and comprehend the correlation between items, without any prior knowledge about the latent structural properties of the instrument, then EFA is the appropriate method to use.

In this Section, a brief description of the descriptive analysis that should be conducted before EFA will be first introduced. Then selection criteria for the number of factors will be presented. Afterwards, the EFA model will be presented for a single selected day, weekly item average and multilevel split approach for the daily measurements across the week. Finally, the rotation method and goodness of fit measures will also be described.

### 4.2 Descriptive measures for exploratory factor analysis model

#### 4.2.1 Kaiser Meyer Olkin

Before proceeding in conducting EFA, an important step is to first assess whether such a method is appropriate to employ. This means that it is important to check whether



groups of variables are related due to some unobserved factor, and whether the strength of that relationship is strong enough to necessitate its evaluation through a model.

A statistical measure that allows the evaluation of EFA, is called Kaiser-Meyer-Olkin (KMO). This measure is a useful tool to assess the suitability of the observed data to conduct EFA. The notion behind this measure is that if we examine the correlation of two variables given that the other variables are partialled out, then the strength of that correlation should be lower. This type of correlation is often called partial correlation coefficients.

KMO generally ranges from 0 to 1, with values close to 1 indicating that the observed data are appropriate for EFA and values close to 0 showing that there is no need for conducting EFA. Based on equation 4.1 it is clear that the correlation of the variables should be high, and the partial correlation should be low, so when  $\sum \sum_{i \neq j} a_{ij}^2$  is close to 0 in an ideal scenario then KMO will be close to 1. A conventional cut-off value for KMO is 0.5, with values greater than 0.5 indicating that the dataset is appropriate for EFA.

$$KMO = \frac{\sum \sum_{i \neq j} r_{ij}^2}{\sum \sum_{i \neq j} a_{ij}^2 + \sum \sum_{i \neq j} r_{ij}^2} \quad (4.1)$$

where  $r_{ij}$  and  $a_{ij}$  are the sample correlation and partial correlation coefficients, respectively.

#### 4.2.2 Measure of sampling adequacy

A similar measure with KMO serving the same purpose is the Measure of Sampling Adequacy (MSA). The formula is similar with KMO but in this case this measure focuses on the adequacy of each variable separately whereas KMO focuses on the overall adequacy of the whole dataset (including all variables). Similarly, values close to 1 indicate that the variable is appropriate for using in EFA and values close to 0 indicate that the variable should be excluded from the analysis. The cut-off values are the same as in KMO.

$$MSA_i = \frac{\sum_j r_{ij}^2}{\sum_j a_{ij}^2 + \sum_j r_{ij}^2} \quad (4.2)$$

where  $r_{ij}$  and  $a_{ij}$  are the sample correlation and partial correlation coefficients, respectively.



## 4.3 Selection criterion for the number of factors

### 4.3.1 Kaiser criterion

When it comes to EFA, one of the most crucial steps is to determine the number of factors to retain. That is why a wide array of methods has been developed to serve that purpose. One of the most well known methods is the Kaiser criterion (Kaiser, 1960). This method is commonly used in EFA (Garrido, Abad, & Ponsoda, 2013) and it is based on the idea that if variables are uncorrelated then the correlation matrix will be an identity matrix (i.e., the eigenvalues of that matrix would be equal to 1). Consequently, in an observed correlation matrix where high correlation would be expected between items within each factor, it is expected that eigenvalues would be greater than 1. So this method suggests that factors with eigenvalues greater than 1 to be retained.

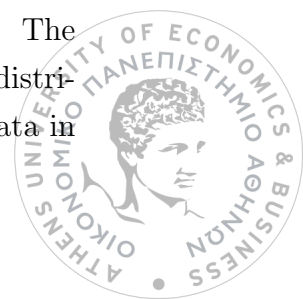
This method was introduced by Kaiser in 1960, who proposed that eigenvalues greater than 1 implies that the corresponding factor explains more variance than a single variable, so it should be retained.

Such a method can be visualized through scree plot (Cattell, 1966), which is a graph that shows the eigenvalues of the correlation matrix on the y-axis and the number of factors on the x-axis. The eigenvalues are in descending order, and the graph shows how eigenvalues decrease when the number of factors increases.

Although such a method is really popular, it has its limitations. It is mistakenly assumed by many researchers that Kaiser criterion provides the actual number of factors when in reality it provides just a lower bound for the number of factor to be retained. However, even in cases where the lower bound of the number of factor is determined, there is still an issue. This is because the lower bound that was found based on Kaiser criterion it is based on population correlation matrix rather than the observed correlation matrix. This has a serious complication, as Kaiser criterion based on the sample will either overestimate or underestimate the number of factors (Cattell & Vogelmann, 1977; Cliff, 1988; Horn, 1965; Browne, 1968). Too many factors will lead to a complex solution with no interpretability, while too few factors will mask important factors, leading to misleading or incomplete results.

### 4.3.2 Empirical Kaiser criterion

Another approach for factor identification is the empirical Kaiser criterion (Braeken & Van Assen, 2017). This approach takes advantage of the sampling distribution of eigenvalues under the null hypothesis that the observed data are uncorrelated. The distribution of the eigenvalues under that assumption is the Marchenko-Pastur distribution, and it is used to compare the estimated eigenvalues from the observed data in



comparison with what we would expect if data were randomly drawn. This is achieved by using a reference value of Marchenko-Pastur distribution (Marchenko & Pastur, 1967). By assuming that  $J$  is the number of variables,  $n$  is the sample size and  $\gamma = \frac{J}{n}$ , the density function of the sampling distribution of eigenvalues  $L = [l_1, l_2, \dots, l_J]$  under the null hypothesis that the data are uncorrelated is described below:

$$d(l) = \begin{cases} \frac{\sqrt{(l_{up}-l)(l-l_{low})}}{2\pi\gamma l} & l_{low} \leq l \leq l_{up} \\ 0 & \text{otherwise} \end{cases}$$

Additionally, if  $\gamma$  is greater than 1 then the density function is equal to  $1 - \frac{1}{\gamma}$  where  $\gamma = \frac{J}{n}$

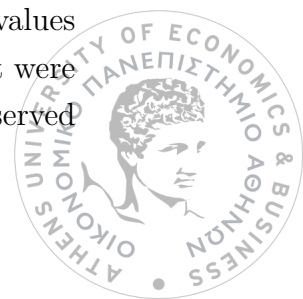
This method also accounts for the relationship between consecutive eigenvalues, as the first eigenvalue will always explain the highest proportion of variability in a dataset, with iterative decreases for each eigenvalue in sequence. So, inference regarding the number of factors to retain via the empirical Kaiser criterion is based on the number of observed eigenvalues that are both greater than 1 and greater than the reference eigenvalues based on the sampling distribution of eigenvalues under the null model. The reference values are described in equation 4.3.

$$l_j^{EKC} = \max\left(\frac{J - \sum_{j=0}^{j-1} l_j}{J - j + 1} (1 + \sqrt{\gamma})^2, 1\right) \quad (4.3)$$

Although this method takes into account the sampling distribution of eigenvalues, in contrast with Kaiser criterion which is based on the population level rather than the sampling level, it comes with some limitations. There has not been extensive research on the impact of number of variables per factors, cross loading and heterogeneous factor loadings in the performance of empirical Kaiser criterion (Braeken & Van Assen, 2017).

### 4.3.3 Parallel analysis

An alternative approach, which also focuses on the sample level of the data rather than population level, is parallel analysis. Parallel analysis is considered among the best methods to identify the number of factors (Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1982) although it is not as widely used as the Kaiser criterion. Parallel analysis tries to take into account that there is a sampling error in the data and that the population correlation matrix is unknown. The method is implemented by taking many random samples, calculating the eigenvalues of these sample correlation matrices, and then computing the average of those eigenvalues for comparison with the observed eigenvalues. In this approach, the ideal scenario would be if the observed eigenvalues exceed the average eigenvalues of the correlation matrices based on the data that were drawn randomly. So the number of factors is determined by how much of the observed



eigenvalues exceed the corresponding eigenvalues from the uncorrelated data.

Although averaging the values of the eigenvalues across many random samples is a common approach, 95<sup>th</sup> percentile could be also computed. These two could provide similar results, but when this is not the case, 95<sup>th</sup> percentile is preferred as it is more robust to the problem of overfactoring (Horn, 1965).

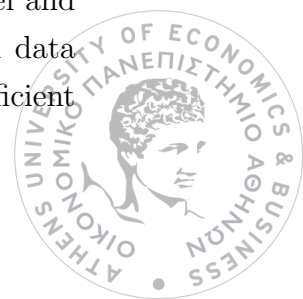
While Kaiser criterion is based on a specific threshold which will determine how many factors to retain, parallel analysis is a more rigorous approach as it compares the observed eigenvalues with the eigenvalues drawn from uncorrelated data. Generally, parallel analysis is a more reliable tool for large datasets. However, it is computationally heavier than Kaiser criterion, so it requires more time. Parallel analysis also seems to underperform when the sample size is small, as it requires the analyst to randomly generate data. This means that when there is a large sample size, there is more information available. Thus, the eigenvalues based on the simulated data are more representative of the true eigenvalues under the assumption that the data are randomly drawn. Parallel analysis also seems to underperform when the factor structure of the data is unidimensional and when the factors are highly correlated, especially when within each factor there are few variables (Buja & Eyuboglu, 1992; Dinno, 2009).

## 4.4 Exploratory ordinal factor analysis for single selected say approach

The current work emphasizes on various correlation matrices due to the use of diverse data handling approaches. As previously mentioned the first approach, which is under investigation, corresponds to the single selected day which utilizes a proportion of the data, the second approach corresponds to the weekly item average approach which aggregates the data, and the last approach corresponds to the within- and between-individual analysis that utilize all the data. Such methods require a careful consideration before proceeding to EFA, as the input correlation matrix is based on the nature of the data that is being utilized.

### 4.4.1 Input correlation matrix: polychoric correlation

When data are on an ordinal scale, as in the case of the simulation study, the single selected day approach will result in data on an ordinal scale. This is an important consideration as it affects the input correlation matrix, the selection of the model and the estimation method. A common approach to study the correlation between data that are on an ordinal scale, is to use polychoric and tetrachoric correlation coefficient



while Pearson correlation coefficient, is less appropriate. Pearson correlation is not usually preferred for such data as it underestimates the true correlation as all individuals that are suited at the same interval belong to the same category, which results in the reduction of the variability of the data.

Polychoric correlation is used for ordinal data with more than 2 categories, and tetrachoric correlation is used for dichotomous data. They can range from -1 to 1 and high positive values indicate strong positive association between the ordered data, strong negative values indicate that there is strong negative association between the ordered data, while values close to zero indicate that there is no association at all. These type of correlations are estimated by assuming that the observed ordinal data are a manifestation of some underlying continuous variables. Let's assume that  $X$  and  $Y$  are two items with ordinal scale and  $k_1$  and  $k_2$  are the number of categories, respectively. Let's also assume that  $X^*$  and  $Y^*$  are two continuous latent variables, which follow a normal distribution. Each of the latent continuous variables is categorized into  $k_1$ ,  $k_2$  categories based on  $k_1-1$ ,  $k_2-1$  thresholds. The joint distribution of the latent continuous variables is assumed to be bi-variate normal and the correlation of those two variables is the polychoric correlation when  $k_1, k_2 > 2$  and tetrachoric when  $k_1 = k_2 = 2$ . Although there is a strong assumption of bi-variate normality across the two latent variables, even in the case where such an assumption is violated, it can still give robust results (Coenders, Satorra, & Saris, 1997). The observed responses can be considered as a manifestation of the distribution of the latent responses as was previously flagged. More specifically for the general case where  $k_1, k_2 > 2$   $X_{ij}$  is equal to category  $c_1$  if  $a_{c_1} < X_{ic_1}^* < a_{c_1+1}$  and  $Y_{ij}$  is equal to category  $c_2$  if  $b_{c_2} < Y_{ic_2}^* < b_{c_2+1}$  with  $c_1 = 1, \dots, k_1$  and  $c_2 = 1, \dots, k_2$ .

where:

$i$  : Subscript of the observation

$j$  : Subscript of the category

These thresholds which can be shown as an example in Figure 4.1 are the cut-off points for the marginal distribution of each latent variable, and they are practically the quantiles of the corresponding distributions. They play a crucial role in determining the joint distribution of the latent variables and the polychoric correlation.



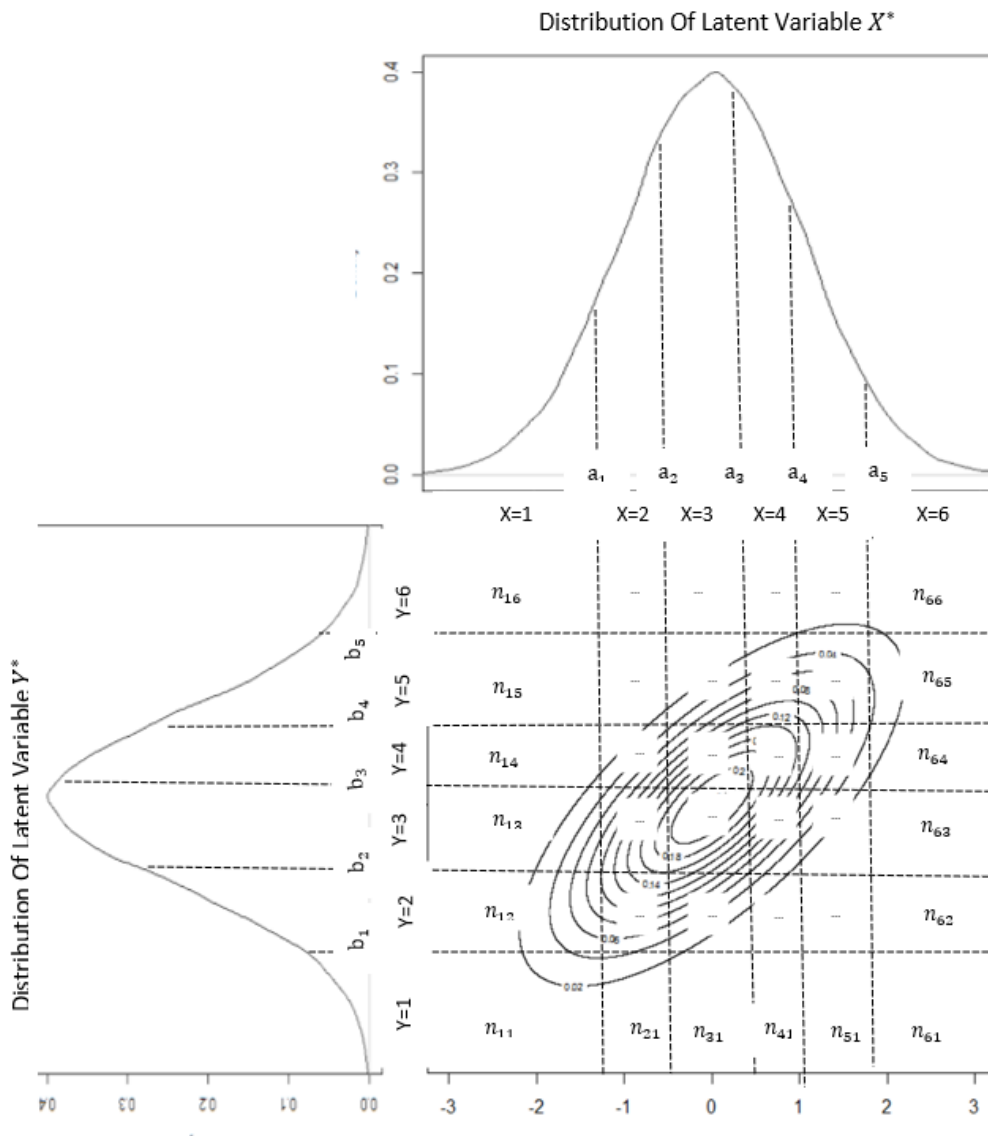


Figure 4.1: Bi-variate and marginal latent response distributions for polytomous items X and Y with 6 categories.

To provide a simple example for the importance of those threshold values, we will assume a simple case where there are two dichotomous items ( $k_1=2, k_2=2$ ). For each item, there will be just one threshold which will determine whether each variable will be suited in category 1 or 2. If  $a_1$  and  $b_1$  are the thresholds of the dichotomous items X and Y, they can be considered as the Z scores of the normal distribution of the latent variables. Those Z score are cut-off points to determine each category of the items. The corresponding probability of observing a latent continuous value which is less than the corresponding cut-off value of the corresponding item can be considered as the probability of observing category 1. Similarly, the probability of observing a latent continuous value higher than the corresponding cut-off value can be considered as the probability of observing category 2.

For this thesis the focus will be on polychoric correlation as the items on the simulation study are on an ordinal scale (0-4). Such a correlation as previously discussed tries to employ information of some unobserved variables, whereas for the observed values, the only source of information is the frequency of observing each combination of the item levels across the items. If for example we have only two items, variable  $X$  and  $Y$  which have  $k_1$  and  $k_2$  categories respectively, they will result in a  $k_1 \times k_2$  contingency table. This table is showed in Table 4.1 and  $n_{ij}$  are the observed frequencies of the ordered data with  $i=1, \dots, k_1$  and  $j=1, \dots, k_2$  and  $\pi_{ij}$  is the probability for an observation to fall into the cell  $i$  and  $j$ . In order to estimate the thresholds and polychoric correlation, we need to maximize the likelihood of the multinomial distribution of the contingency table.

Table 4.1:  $k_1 \times k_2$  contingency table of  $X$  and  $Y$  with  $k_1, k_2$  categories respectively

		Y				
$X$	$n_{11}$	$n_{12}$	$n_{13}$	$\dots$	$n_{1k_2}$	
	$n_{21}$	$n_{22}$	$n_{23}$	$\dots$	$n_{2k_2}$	
	$n_{31}$	$n_{32}$	$n_{33}$	$\dots$	$n_{3k_2}$	
	$n_{41}$	$n_{42}$	$n_{43}$	$\dots$	$n_{4k_2}$	
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	
	$n_{k_1 1}$	$n_{k_1 2}$	$n_{k_1 3}$	$\dots$	$n_{k_1 k_2}$	

The likelihood and log-likelihood of the multinomial distribution are described in equation:

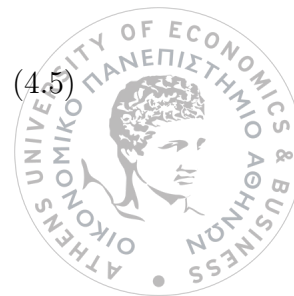
$$L = P(N_{11} = n_{11}, N_{12} = n_{12}, \dots, N_{k_1 k_2} = n_{k_1 k_2}) = \frac{N!}{n_{11}! n_{12}! \dots n_{k_1 k_2}!} \prod_{i=1}^{k_1} \prod_{j=1}^{k_2} \pi_{ij}^{n_{ij}}$$

$$l = \log C + \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} n_{ij} \log \pi_{ij} = 1 \tag{4.4}$$

where  $C$  is equal to  $\frac{N!}{n_{11}! n_{12}! \dots n_{k_1 k_2}!}$

As was previously flagged  $\pi_{ij}$  is the probability for an observation to fall into category  $i$  and  $j$ . Equivalently, this could be described by the below equation:

$$\begin{aligned} \pi_{ij} &= P(a_{i-1} < X^* < a_i, b_{j-1} < Y^* < b_j) \\ &= \Phi(a_i, b_j) - \Phi(a_i, b_{j-1}) - \Phi(a_{i-1}, b_{j-1}) + \Phi(a_{i-1}, b_j) \end{aligned} \tag{4.5}$$



where  $\Phi$  is the bivariate normal cumulative distribution function of the latent variables  $X^*$  and  $Y^*$  with polychoric correlation  $\rho$ .

This maximum likelihood estimation (MLE) of the contingency table was firstly introduced for a 2-dimensional contingency table (Tallis, 1962; Olsson, 1979), where thresholds parameters and polychoric correlation were estimated simultaneously. Then, the estimation of polychoric correlation from a 2-dimensional contingency table was extended to  $r$  dimensional contingency tables (S.-Y. Lee & Poon, 1985). The proposed methodology was computationally challenging as the number of items increased. That is because when the number of items increase, the number of integration increases as well. So, for instance, for the polychoric correlation among 15 items, a 15-dimensional integration is required. This is the reason a two-step procedure was suggested (Hamdan & Martinson, 1971), where thresholds are estimated first through the marginal distributions of the latent variables and then based on those thresholds which are fixed, polychoric  $\rho$  is estimated. This method has as an advantage that it is less computationally heavier than the previous method but estimating correlations independently can result in a non-positive definite matrix (X.-Y. Song & Lee, 2003) which can be an important issue when for instance someone is interested in proceeding to factor analysis.

In the first approach the estimates are produced based on the below first derivatives of the multinomial distribution where both thresholds and estimated polychoric correlation are simultaneously:

$$\begin{aligned} \frac{\partial l}{\partial \rho} &= \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \rho} \\ &= \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{n_{ij}}{\pi_{ij}} (\phi(a_i, b_j) - \phi(a_{i-1}, b_j) - \phi(a_i, b_{j-1}) + \phi(a_{i-1}, b_{j-1})) \end{aligned} \quad (4.6)$$

$$\begin{aligned} \frac{\partial l}{\partial a_k} &= \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial a_k} \\ &= \sum_{j=1}^{k_2} \left( \frac{n_{kj}}{\pi_{kj}} - \frac{n_{k+1j}}{\pi_{k+1j}} \right) \left( \frac{\partial \Phi(a_k, b_j)}{\partial a_k} - \frac{\partial \Phi(a_k, b_{j-1})}{\partial a_k} \right) \end{aligned} \quad (4.7)$$



$$\begin{aligned} \frac{\partial l}{\partial b_n} &= \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial b_n} \\ &= \sum_{i=1}^{k_1} \left( \frac{n_{in}}{\pi_{in}} - \frac{n_{in+1}}{\pi_{in+1}} \right) \left( \frac{\partial \Phi(a_i, b_n)}{\partial b_n} - \frac{\partial \Phi(a_{i-1}, b_n)}{\partial b_n} \right) \end{aligned} \quad (4.8)$$

In the second approach the thresholds are estimated first through the marginal distribution of the latent continuous variables and then based on the fixed values of the thresholds, polychoric correlation is estimated.

The threshold parameters are calculated by the two below equations:

$$a_i = \Phi^{-1}(P_{.i}) \quad (4.9)$$

$$b_j = \Phi^{-1}(P_{.j}) \quad (4.10)$$

where  $P_{.i} = \sum_{k=1}^i \sum_{j=1}^{k_1} P_{kj}$  and  $P_{.j} = \sum_{i=1}^{k_2} \sum_{k=1}^j P_{ik}$

Then polychoric correlation is calculated based on the below derivative by using the estimated thresholds in the previous stage:

$$\begin{aligned} \frac{\partial l}{\partial \rho} &= \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{n_{ij}}{\pi_{ij}} \frac{\partial \pi_{ij}}{\partial \rho} \\ &= \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{n_{ij}}{\pi_{ij}} (\phi(a_i, b_j) - \phi(a_{i-1}, b_j) - \phi(a_i, b_{j-1}) + \phi(a_{i-1}, b_{j-1})) \end{aligned} \quad (4.11)$$

#### 4.4.2 Model

Given that the selection of a single day will not modify the ordered scale of the data, the ordinal EFA model could be employed by using as input matrix polychoric correlation matrix. Let  $y_{ti}$  be a  $p$ -dimensional vector, including the observed variables at time  $t$  for individual  $i$ .

The form of the model is described as follows:

$$y_{ti}^* = L^* \eta_i^* + \epsilon_{ti}^* \quad (4.12)$$



where:

$y_{ti}^*$  :  $p$ -dimensional vector of the unobserved latent continuous variables at day  $t$  for individual  $i$

$\eta_i^*$  :  $k \times 1$  vector of the factor scores of individual  $i$

$L^*$  :  $p \times k$  factor loading matrix

$p$  : Number of variables

$k$  : Number of factors

Although this form of the model is adequate to describe the EFA model, an alternative way is to describe it through the covariance matrix. This is a more natural way to describe such a model, as the main goal of a factor analysis model is to reproduce the covariance matrix of the observed data. The way it could be reproduced, is described in equation 4.13

$$\Sigma^* = L^* Cov(\eta) L^{*'} + \Psi^* \quad (4.13)$$

where:

$\Sigma^*$  :  $p \times p$  covariance matrix of the latent continuous response options

$L^*$  :  $p \times k$  factor loading matrix

$Cov(\eta)$  :  $k \times k$  Correlation among factors

$L^* Cov(\eta) L^{*'}$  : Communality

$\Psi^*$  :  $p \times p$  Specificity

$p$  : Number of variables

$k$  : Number of factors

Based on the above definitions, communality is the amount of variance of the unobserved latent continuous responses that the factors explain and specificity is the amount of variance of the unobserved latent continuous responses that it is not explained through the factors.

The above model is identified based on general guidance of identification rules (Chen, Bollen, Paxton, Curran, & Kirby, 2001) but the variances of unobserved latent responses are not identified because  $y^*$  is unobserved. This the reason for which in ordinal factor analysis the model can be identified by setting  $\Psi^*$  as described below:



$$\Psi^* = I - \text{diag}(L^{*'} \text{Cov}(F) L^*) \quad (4.14)$$

Note that this model assumes that there is correlation among factors, so this model is non-orthogonal, whereas orthogonal models assume that there is no correlation among factors.

### 4.4.3 Distributional assumptions

The assumptions of the EFA for the single selected day are the following:

- $E(\eta^*)=0$
- $E(\epsilon^*)=0$
- $\text{Cov}(\epsilon^*)=\Psi^*$
- $\text{Cov}(\epsilon_{iz}^*, \eta_{jl}^*)=0 \forall i \neq j$
- $y_i^*$  follows multivariate normal distribution

where:

$\Psi^*$  : Diagonal matrix with  $\psi_1^*, \psi_2^*, \dots, \psi_p^*$  being the diagonal values

$\eta_{jz}^*$  : Factor score for the  $j$  individual and  $z$  factor

$\epsilon_{il}^*$  : Residual for the  $i$  individual and  $l$  variable

$\epsilon^*$  :  $p$ -dimensional vector with the residuals

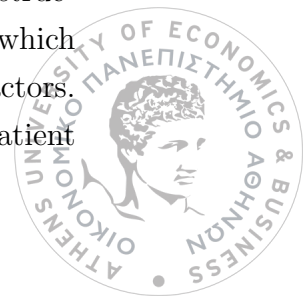
$\eta^*$  :  $k$ -dimensional vector of the factor score

$p$  : Number of variables

$k$  : Number of factors

### 4.4.4 Additional assumptions

Apart from the distributional assumptions, there are also some additional assumptions associated with EFA. These assumptions are relevant to the current modelling approaches that are adopted for analysing daily diary data. More specifically, the single selected day approach which is presented in this Chapter necessitates the factor structure to remain the same across time, including the number of factors, the item which belong to each factor and the association strength between the items and the factors. Those assumptions are quite important and if they are violated then when a patient

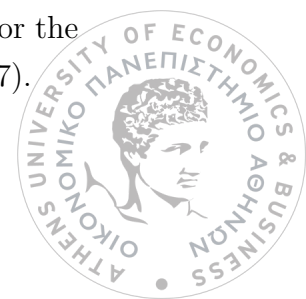


change his response score across time, the observed change is attributed to the fact that the construct has changed and not due to actual change. So the patient interpret differently the instrument across time. This practically means that when someone is using the single selected day approach and select a specific day, it is necessary that the participants of the study interpret the same way the items across time, otherwise the results can be completely different based on which day is being selected. In addition to this, for such a method, within-individual variability is considered a statistical nuisance. So if someone would implement EFA across different days for a one-week period, the results would be similar. These assumptions are pivotal for the single selected day and if these assumptions do not hold then the results would not be accurate. For greater readability, the assumptions are summarized as follows:

- Scalar invariance: The number of factors, the items that load to each factor does not change across time, the loadings, and the threshold parameters of the factor model do not change across time
- Within-individual variability is a statistical nuisance, so each day will produce theoretically similar results in the factor analysis procedure

#### 4.4.5 Estimation method

Given that the data are ordinal in the single selected day approach and not continuous, MLE is not an optimum method to use. That is because MLE is based on Pearson correlation and not on polychoric correlation. When data are categorical in general, they are inherently considered non-normal (B. Muthén & Kaplan, 1985). Normality of the data is an important assumption of MLE method and its violation may lead to biased Chi-square and standard error estimates, but the parameter estimates will still remain accurate (Bollen, 1989). This also leads to biased results in the goodness of fit measures that are a function of Chi-square statistic, such as RMSEA or TLI. This also applies even in cases where the model is correctly specified but non normality occurs. In the case where the data are non-normal and ordinal in particular, standard error and TLI are underestimated and Chi-square and RMSEA are inflated when using MLE (Mindrila, 2010). So ordinal data shouldn't be handled as continuous. The only case where ordinal data can be treated as continuous is when item levels are at least five and there is little violation to normality (Schumacker & Beyerlein, 2000). Thus, as the number of categories decreases, so does the bias of standard error and Chi square statistic (Mindrila, 2010). This bias is even greater when the sample size or the correlation of items and factors is small (Babakus, Ferguson Jr, & Jöreskog, 1987).



Multivariate normality is an important assumption when conducting factor analysis, and it is usually quantified based on the skewness and kurtosis of the data. Skewness is a measure of quantifying the lack of symmetry of the data, and values greater than 1 and less than -1 indicate that the data are highly skewed. Kurtosis is a measure/quantification of the extent to which the tails are heavy or light compared to the normal distribution. Leptokurtic data indicates that the data are gathered at a high probability in the centre of the distribution, whereas platykurtic data indicates that the data are gathered mostly in the area of right and left tail of the distribution. Both skewness and kurtosis play a crucial role in obtaining accurate results in the statistical process that is implemented (Schumacker & Lomax, 2004), including factor analysis. For example, if the data are leptokurtic, the amount of bias in standard error is higher (Hoogland & Boomsma, 1998).

Consequently, when non normality occurs and the number of item levels is small, more preferable estimation methods are Weighted Least Squares (WLS) (B. Muthén, 1978, 1984), DWLS (B. O. Muthén, 1997) and Unweighted Least Squares (ULS) (B. O. Muthén, 1993). The model is more specifically estimated in 3 stages (note that the first two stages are also described in the estimation of polychoric correlation). In the first stage, threshold parameters are estimated for each variable separately by using maximum likelihood method. In the second stage, polychoric correlation of each pair of variables is estimated based on the estimated thresholds produced in the previous stage. In the third stage, the polychoric correlations of the model are estimated by using the estimated polychoric correlation of the two previous stages.

These parameters in the third stage are estimated by minimizing the least square function. In the case where there is no restriction in the threshold parameters, the least square function based on polychoric correlation can be used (B. Muthén, 1978) as follows:

$$F = (\hat{\rho} - \rho(\theta))' \widehat{W} (\hat{\rho} - \rho(\theta)) \quad (4.15)$$

- For the WLS estimation  $\widehat{W} = \widehat{\Gamma}^{-1}$
- For the DWLS estimation  $\widehat{W} = \text{diag}(\widehat{\Gamma})^{-\frac{1}{2}}$
- For the ULS estimation  $\widehat{W} = I$  (identity matrix)



where:

$\widehat{\Gamma}$  : Asymptotic covariance matrix of the estimated polychoric correlation

$\widehat{\rho}$  : Estimated polychoric correlation

$\rho(\theta)$  : Model based polychoric correlation

DLWS and ULS and their variants (e.g., Weighted Least Square mean and variance adjusted (WLSMV), Unweighted Least Square mean and variance adjusted) are generally more preferable than WLS for categorical data with small sample size.  $W$  is usually unstable, especially in small sample sizes. This instability based on the literature has been proved to result in misleading fit statistics (Dolan, 1994; Flora & Curran, 2004), and biased standard errors (B. Muthén & Kaplan, 1985). Generally WLS is advised to be used only when the sample size is at least 1,000 (Hoogland & Boomsma, 1998) and based on the complexity of the model and the data, it can even require a sample size greater than 4000 (Boomsma & Hoogland, 2001). Although both ULS and DWLS can be used for categorical, for this thesis DWLS will be employed, which, although it might give less biased estimates it can have less convergence issues under small sample sizes (Forero, Maydeu-Olivares, & Gallardo-Pujol, 2009).

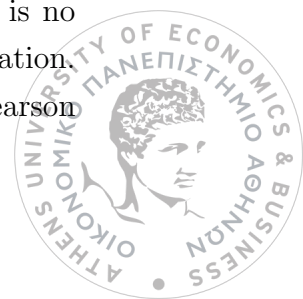
## 4.5 Exploratory factor analysis for item average approach

### 4.5.1 Input correlation matrix: Pearson correlation matrix

The second approach refers to averaging the item across time for each individual, and it is referred to as the item average approach. Based on the simulation study, the data are on an ordinal scale, so averaging the items across time will result in transforming the data from an ordinal to continuous scale, so an appropriate correlation matrix to describe the relationship of the items for this type of scale is Pearson correlation matrix (Cohen et al., 2009).

Pearson correlation is used to assess the linear correlation between two continuous variables, and for this simulation study the main focus is the pairwise linear correlation between the items of a questionnaire. This type of correlation ranges from -1 to 1 and if it is 1 then the items have a strong positive association, if it is 0 then there is no linear association at all and if it is -1 then the items have a strong negative association.

By assuming that  $X$  and  $Y$  are two random variables for the population, Pearson



correlation is:

$$\begin{aligned}\rho_{X,Y} &= \frac{Cov(X,Y)}{\sigma_X\sigma_Y} \\ &= \frac{E[XY] - E[X]E[Y]}{\sqrt{E[X^2] - E[X]^2}\sqrt{E[Y^2] - E[Y]^2}}\end{aligned}\quad (4.16)$$

where:

$Cov(X,Y)$  : Population covariance matrix between variable  $X$  and variable  $Y$

$\sigma_X$  : Population standard deviations of variable  $X$

$\sigma_Y$  : Population standard deviations of variable  $Y$

$E$  : Expected value

The population correlation matrix will be:

$$Cor_{Population} = \begin{vmatrix} 1 & \rho_{X_1X_2} & \rho_{X_1X_3} & \rho_{X_1X_4} & \cdots & \rho_{X_1X_p} \\ \rho_{X_2X_1} & 1 & \rho_{X_2X_3} & \rho_{X_2X_4} & \cdots & \rho_{X_2X_p} \\ \rho_{X_3X_1} & \rho_{X_3X_2} & 1 & \rho_{X_3X_4} & \cdots & \rho_{X_3X_p} \\ \rho_{X_4X_1} & \rho_{X_4X_2} & \rho_{X_4X_3} & 1 & \cdots & \rho_{X_4X_p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \rho_{X_pX_1} & \rho_{X_pX_2} & \rho_{X_pX_3} & \rho_{X_pX_4} & \cdots & 1 \end{vmatrix}\quad (4.17)$$

where:

$\rho_{X_iX_j}$  : Pearson correlation between variable  $X_i$  and  $X_j$

$i : 1, \dots, p$

$j : 1, \dots, p$

Assuming that  $n$  is the number of observation drawn from a sample and that  $(x_1, y_1), \dots, (x_n, y_n)$  are the paired data, the sample Pearson correlation is:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}\quad (4.18)$$



where:

$N$  : Sample size

$x_i, y_i$  : Sample points for individual  $i$

The corresponding sample correlation matrix for the paired data consisting  $p$  variables is :

$$Cor_{Sample} = \begin{pmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} & r_{x_1x_4} & \cdots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & r_{x_2x_3} & r_{x_2x_4} & \cdots & r_{x_2x_p} \\ r_{x_3x_1} & r_{x_3x_2} & 1 & r_{x_3x_4} & \cdots & r_{x_3x_p} \\ r_{x_4x_1} & r_{x_4x_2} & r_{x_4x_3} & 1 & \cdots & r_{x_4x_p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ r_{x_px_1} & r_{x_px_2} & r_{x_px_3} & r_{x_px_4} & \cdots & 1 \end{pmatrix} \quad (4.19)$$

where:

$r_{x_ix_j}$  : Sample Pearson correlation between variables

$x_i, y_i$  : Paired data that are drawn from a random sample

$i : i, \dots, p$

$j : j, \dots, p$

This measure is widely used to assess the relationship between continuous variables, but it has some assumptions. Firstly, it assumes that there is a linear relationship between the two variables that are on continuous scale. Secondly, the data are drawn from a random sample that follows normal distribution. Lastly, there is no outlier in the data as it can significantly impact the results. In case there is a non-linear relationship between the continuous data and Pearson correlation is used, then the measure will not correctly reflect the true association strength between the variables of interest. As for the violation of normality, it will not affect the measure itself, but it will affect the hypothesis testing for assessing the statistical significance of the correlation. Finally, outliers may weaken or strengthen the association strength of the variables, and lead to the misrepresentation of the association strength of the variables, which could lead to misleading results.



## 4.5.2 Model

Given that the average weekly item approach transforms the ordered data to continuous, a common EFA model could be employed by using as input correlation matrix Pearson correlation matrix. Let  $y_{ti}$  be a  $p$ -dimensional vector, including the observed variables at time  $t$  for individual  $i$ .

The form of the model is described as follows:

$$\bar{y}_i = L\eta_i + \epsilon_i \quad (4.20)$$

where:

$\bar{y}_i$  :  $p$ -dimensional vector of the average value

of observed variables across a study period for individual  $i$

$\eta_i$  :  $k \times 1$  vector of the factor scores of individual  $i$

$\epsilon_i$  :  $p$ -dimensional vector of the residuals for individual  $i$

$L$  :  $p \times k$  factor loading matrix

$p$  : Number of variables

$k$  : Number of factors

Although this form of the model is adequate to describe the EFA, an alternative way is to describe it through the covariance matrix. This is a more natural way to describe such a model, as the main goal of a factor analysis model is to reproduce the covariance matrix of the observed data. The way it could be reproduced is described in equation 4.21.

$$\Sigma = LCov(\eta)L' + \Psi \quad (4.21)$$



where:

$\Sigma$  :  $p \times p$  covariance matrix of the observed data

$L$  :  $p \times k$  factor loading matrix

$Cov(\eta)$  :  $k \times k$  correlation among the factors

$LCov(\eta)L'$  : Communalities

$\Psi$  :  $p \times p$  specificity

$p$  : Number of variables

$k$  : Number of factors

Note that this model assumes that there is correlation among factors, so this model is non-orthogonal, whereas orthogonal models assumes that there is no correlation among factors.

### 4.5.3 Distributional assumptions

The assumptions of the EFA for the item average approach are the following:

- $E(\eta)=0$
- $E(\epsilon)=0$
- $Cov(\epsilon)=\Psi$
- $Cov(\epsilon_{il}, \eta_{jz})=0 \forall i \neq j$
- $\bar{y}_i$  follows a multivariate normal distribution

where:

$\Psi$  : Diagonal matrix with  $\psi_1, \psi_2, \dots, \psi_p$  being the diagonal values

$\eta_{jz}$  : Factor score for the  $j$  individual and  $z$  factor

$\epsilon_{il}$  : Residual for  $i$  individual and  $l$  variable

$\epsilon$  :  $p$ -dimensional vector for residuals

$\eta$  :  $k$ -dimensional vector of the factor score

$p$  : Number of variables

$k$  : Number of factors



#### 4.5.4 Additional assumptions

Similarly, with the single selected day approach, weekly item average approach require the following assumptions:

- Scalar invariance: The number of factors, the items that load to each factor does not change across time, the loadings, and the threshold parameters of the factor model do not change across the time period that the items are average.
- Within-individual variability is a statistical nuisance and between-individual variability explains the higher proportion of the variance of the data across the time interval in which the observations are averaged

#### 4.5.5 Estimation method

For continuous data, an appropriate estimation method to estimate the parameters of EFA model is to use MLE. The estimated parameters for the EFA are loadings, error variances and factor variances and covariances given that the model is non-orthogonal.

The equation 4.22 describes the probability of observing the sample covariance matrix given the estimated parameters in log scale for EFA model with correlated factors. This likelihood is maximized with respect to the estimated loadings, error variances and factor variances and covariances.

$$F_{ML} = L(\bar{y}_{i.}, L, \Psi, \eta) = -\frac{G}{2} [p \log(2\pi) + \log |LCov(\eta)L' + \Psi| + tr((LCov(\eta)L' + \Psi)^{-1}S)] \quad (4.22)$$

where:

$Cov(\eta)$  :  $k \times k$  covariance matrix of the factors

$L$  :  $p \times k$  covariance matrix of the factors

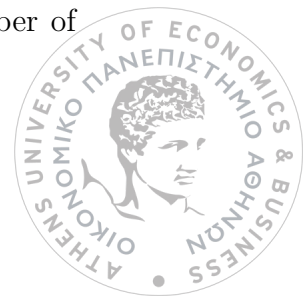
$S$  :  $p \times p$  sample covariance matrix

$p$  : Number of variables

$k$  : Number of factors

$G$  : Number of individuals

Note that with this estimation method, there is a restriction regarding the number of factors that can be obtained. If the number of variables is  $p$ , then the number of factors that can be obtained is the integer part of  $\frac{p}{2}$ .



## 4.6 Exploratory factor analysis for the split multi-level approach

The split approach is referring to the decomposition of the total covariance matrix to within- and between-individual covariance matrix. So in the exploratory framework where the within- and between- individual analysis are conducted separately, the input covariance matrix will be within-individual covariance matrix for the within-individual analysis and between-individual covariance matrix for the between-individual analysis.

### 4.6.1 Input covariance matrix: Within-individual covariance matrix

The input covariance matrix for within-individual analysis will be:

$$S_{pooled} = \frac{\sum_{i=1}^n \sum_{t=1}^{t_i} (y_{ti} - \bar{y}_{.i})(y_{ti} - \bar{y}_{.i})'}{N - G} \quad (4.23)$$

where:

$n$  : Number of individuals

$t_i$  : Number of observations within the individual  $i$

$y_{ti}$  : Observation at time  $t$  for individual  $i$

$\bar{y}_{.i}$  : Average value across time for individual  $i$

### 4.6.2 Input covariance matrix: Between-individual covariance matrix

The input covariance matrix for between-individual analysis will be:

$$S_{between} = \frac{\sum_{i=1}^n (\bar{y}_{..} - \bar{y}_{.i})(\bar{y}_{..} - \bar{y}_{.i})'}{n - 1} \quad (4.24)$$

where:

$n$  : Number of individuals

$\bar{y}_{..}$  : Overall mean across all individuals

$\bar{y}_{.i}$  : Average value across time for individual  $i$



### 4.6.3 Model

In the split approach, all the data are utilized, and common factor analysis is employed by using as input covariance matrix the within-individual covariance matrix in the one case and between-individual covariance matrix in the other. The theory behind the decomposition of the data was mainly built on continuous data, so although the data are on an ordinal scale, common factor analysis under MLE theory will be employed.

#### Within-individual model

Let  $y_{ti}$  be a  $p$ -dimensional vector, including the observed variables at time  $t$  for the individual  $i$ .

The form of the model is described as follows:

$$y_{ti} - \bar{y}_{.i} = L_w \eta_{w_{ti}} + \epsilon_{w_{ti}} \quad (4.25)$$

where:

$\bar{y}_{.i}$  :  $p$ -dimensional vector of the average value of observed variables across time for individual  $i$

$\eta_{w_{ti}}$  :  $k \times 1$  vector of the within-level factor scores of individual  $i$  at time  $t$

$\epsilon_{w_{ti}}$  :  $p$ -dimensional vector of the within-level residuals for individual  $i$  at time  $t$

$L_w$  :  $p \times k$  within-level factor loading matrix

$p$  : Number of variables

$k$  : Number of factors

Alternatively, the model could be described based on the within-individual covariance matrix as follows:

$$\Sigma_w = L_w Cov(\eta_w) L_w' + \Psi_w \quad (4.26)$$



where:

$\Sigma_w$  :  $p \times p$  within-individual covariance matrix of the observed data

$\eta_w$  :  $p \times k$  within-level factor loading matrix

$Cov(\eta_w)$  :  $k \times k$  covariance matrix among the within-level factors

$L_w Cov(\eta_w) L_w'$  : Within-level communality

$\Psi_w$  : Within-level specificity

$p$  : Number of variables

$k$  : Number of factors

### Between-individual model

Let  $y_{ti}$  be a  $p$ -dimensional vector, including the observed variables at time  $t$  for the individual  $i$ .

The form of the model is described as follows:

$$\bar{y}_{.i} = L_b \eta_{b_i} + \epsilon_{b_i} \quad (4.27)$$

where:

$\eta_{b_i}$  :  $k \times 1$  vector of the between-level factor scores of individual  $i$

$L_b$  :  $p \times k$  between-level factor loading matrix

$\epsilon_{b_i}$  :  $p$ -dimensional vector of the between-level residuals for individual  $i$  at time  $t$

$p$  : Number of variables

$k$  : Number of factors

Alternatively, the model could be described based on the between-individual covariance matrix as follows:

$$\Sigma_b = L_b Cov(\eta_b) L_b' + \Psi_b \quad (4.28)$$



where:

$\Sigma_b$  :  $p \times p$  between-individual covariance matrix of the observed data

$\eta_{L_b}$  :  $p \times k$  between-level factor loading matrix

$Cov(\eta_b)$  :  $k \times k$  Covariance matrix among the between-level factors

$L_b Cov(\eta_b) L_b'$  : Between-level communality

$\Psi_b$  :  $p \times p$  between-level specificity

$p$  : Number of variables

$k$  : Number of factors

#### 4.6.4 Distributional assumptions

The assumptions of the EFA for split multilevel approach are the following:

##### Within-individual model

For the within-individual model

- $E(\eta_w) = 0$
- $E(\epsilon_w) = 0$
- $Cov(\epsilon_w) = \Psi_w$
- $Cov(\epsilon_{ilt}, \eta_{w_{jzt}}) = 0 \quad \forall i \neq j$
- $\bar{y}_i$  follows a multivariate normal distribution

where:

$\Psi_w$  : Diagonal matrix with  $\psi_{w_1}, \psi_{w_2}, \dots, \psi_{w_p}$  being the diagonal values

$\eta_{w_{jzt}}$  : Within-level factor score for the  $j$  individual for  $z$  factor at time  $t$

$\epsilon_{w_{ilt}}$  : Within-level residual for  $i$  individual for variable  $l$  at time  $t$

$\epsilon_w$  :  $p$ -dimensional vector for within-level residual

$\eta_w$  :  $k$ -dimensional vector of the within-level factor score

$p$  : Number of variables

$k$  : Number of factors



## Between-individual model

For the between-individual model:

- $E(\eta_b)=0$
- $E(\epsilon_b)=0$
- $Cov(\epsilon_b)=\Psi_b$
- $Cov(\epsilon_{b_{il}}, \eta_{b_{jz}})=0 \forall i \neq j$
- $\bar{y}_i$  follows multivariate normal distribution

where:

$\Psi_b$  : Diagonal matrix with  $\psi_{b_1}, \psi_{b_2}, \dots, \psi_{b_p}$  being the diagonal values

$\eta_{b_{jz}}$  : Between-level factor score for the  $j$  individual for  $z$  factor

$\epsilon_{b_{il}}$  : Between-level residual for  $i$  individual for variable  $l$

$\epsilon_b$  :  $p$ -dimensional vector for between-level residual

$\eta_b$  :  $k$ -dimensional vector of the between-level factor score

$p$  : Number of variables

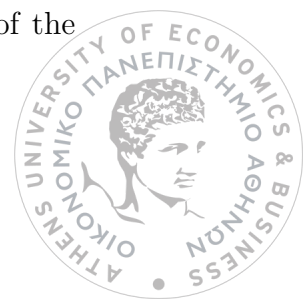
$k$  : Number of factors

### 4.6.5 Additional assumptions

Apart from the distribution assumptions, there are also some additional assumptions. Both within- and between-individual variability are insightful sources of variance. This means that both covariance matrices should explain a significant proportion of the data. Such an assumption is measured through ICC (Koch, 2004).

### 4.6.6 Intraclass Correlation Coefficients

ICC, which was first introduced as a modification measure of Pearson correlation (Fisher, 1954), is generally widely used in social sciences and psychology. It is a statistical tool which is used when there is a hierarchical structure in the data. Generally, it is used to assess the reliability of an instrument, which refers to the precision and validity of its measurement. ICC has different versions, but all serve the same purpose, which is to quantify to what extent the variance of interest explains the total variance of the



observed data. It could be generally defined as follows:

$$\begin{aligned} ICC &= \frac{\text{Variance of interest}}{\text{Total variance}} \\ &= \frac{\text{Variance of interest}}{\text{Variance of interest} + \text{Unwanted variance}} \end{aligned} \quad (4.29)$$

Three common applications of ICC in the area of psychometrics are:

- Test-retest reliability: It is a measure of reliability of an instrument by evaluating its stability across two time points for continuous data
- Inter-rater reliability: It is a measure of the degree of agreement between different raters for evaluating the same concept of interest.
- Intra-rater reliability: It is a measure of the agreement of data measured by 1 rater across two trials

Additionally, ICC serve also other purposes, which include the assessment of appropriateness of multilevel models, such as linear mixed and multilevel factor analysis models.

Let  $y_{ij}$  be an observation for  $i$  group and  $j$  individual and  $X$  is a fixed variable.

The linear mixed model could be described as follows:

$$Y_{ij} = \beta_0 + \beta_1 X + u_i + \epsilon_{ij} \quad (4.30)$$

with  $u_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_b^2)$

So the population ICC for linear mixed models can be defined as follows:

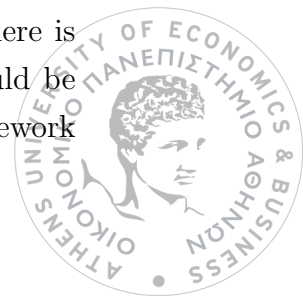
$$\frac{\sigma_b^2}{\sigma_w^2 + \sigma_b^2} \quad (4.31)$$

where:

$\sigma_w^2$  : Within group variance which measures how much variability there is between the observation within the same group

$\sigma_b^2$  : Between group variance which measures how much variability there is between observations across different groups

ICC based on the model can be considered as the correlation of observations across the same cluster. Thus, if data within the same cluster are not related then there is no need to conduct linear mixed model and simple linear regression model could be used. This measure was also introduced under latent variable modelling framework



(B. O. Muthén, 1991) and it is widely used when conducting multilevel factor analysis. As previously mentioned, there are two sources of variances when studying multilevel factor analysis models which is in accordance with linear mixed models: the within- and between-individual variance. However, within-individual variance is usually considered a statistical nuisance when using factor analysis, so it could be considered as the "unwanted variance" based on the general formula of the ICC. On the other hand between-individual variance, could be considered as the "variance of interest".

So, the population ICC for multilevel factor analysis can be described as follows:

$$\begin{aligned} ICC &= \frac{\Sigma_{between}}{\Sigma_{total}} \\ &= \frac{\Sigma_{between}}{\Sigma_{between} + \Sigma_{within}} \end{aligned} \quad (4.32)$$

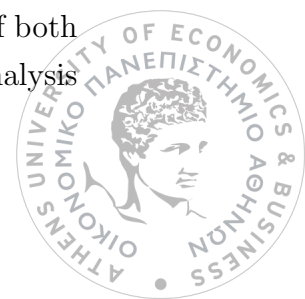
The sample population ICC is described as follows:

$$\begin{aligned} ICC &= \frac{S_{between}}{S_{total}} \\ &= \frac{S_{between}}{S_{between} + S_{within}} \end{aligned} \quad (4.33)$$

where  $S_{between}$  and  $S_{within}$  are described in equation 2.5 and 2.6

ICC is used in multilevel factor analysis models to assess whether both sources of variance explain a significant proportion of the variability of the data. This means that values very close to 0 or 1 will indicate that only one source of variance explains most of the variance of the data. In such cases, multilevel factor analysis is not necessary to conduct. If ICC is very close to 1, this means that the majority of variance is explained by the between-individual differences. Therefore, the data could be considered independent. Similarly, if the ICC value is close to 0 then the majority of variance is explained by within-individual variation. ICC in the multilevel factor analysis framework could be considered as a measure of dependency of the data. In order to assess whether multilevel factor analysis model could be conducted specific cut-off values (Schoemann et al., 2014) could be used:

- If ICC is less than 0.05 then the variance of the data is mostly explained by the within-individual variance, so multilevel factor analysis model is not recommended
- If ICC is between 0.05 and 0.95 then there is substantial contribution of both sources of variance to the total variance of the data, so multilevel factor analysis model is appropriate



- If ICC is greater than 0.95 then the variance of the data is mostly explained by the between-individual variance, so multilevel factor analysis model is not recommended

#### 4.6.7 Estimation method

For implementing the split approach, MLE is utilized.

##### Within-individual model

The estimated parameters for the exploratory within-individual factor analysis model are within-level loadings, within-level error variances and within-level factor variances and covariances given that the model is non-orthogonal.

The equation 4.34 describes the probability of observing the sample within-individual covariance matrix given the estimated parameters in log scale for exploratory within-individual factor analysis model with correlated factors. This likelihood is maximized with respect to the estimated within-level loadings, within-level error variances and within-level factor variances and covariances.

$$\begin{aligned}
 F_{MLE_w} &= L(y_{ti} - \bar{y}_i, L_w, \Psi_w, \eta_w) \\
 &= -\frac{N}{2} [p \log(2\pi) + |\log L_w Cov(\eta_w) L'_w + \Psi_w| \\
 &\quad + tr((L_w Cov(\eta_w) L'_w + \Psi_w)^{-1} S_w)]
 \end{aligned} \tag{4.34}$$

where:

$Cov(\eta_w)$  :  $k \times k$  covariance matrix of within-level factors

$L_w$  :  $p \times k$  within-level factor loading matrix

$S_w$  :  $p \times p$  sample within-level covariance matrix

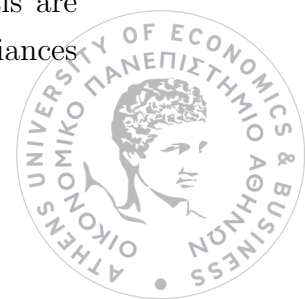
$p$  : Number of variables

$k$  : Number of factors

$N$  : Total number of observations

##### Between-individual model

The estimated parameters for the exploratory between-individual factor analysis are between-level loadings, between-level error variances and between-level factor variances and covariances given that the model is non-orthogonal.



The equation 4.35 describes the probability of observing the sample between-individual covariance matrix given the estimated parameters in log scale for exploratory between-individual factor analysis model with correlated factors. This likelihood is maximized with respect to the estimated between-level loadings, between-level error variances and between-level factor variances and covariances.

$$F_{MLE_b} = L(\bar{y}_i, L_b, \Psi, \eta_b) = -\frac{G}{2} [p \log(2\pi_b) + |\log L_b Cov(\eta_b) L_b' + \Psi_b| + tr((L_b Cov(\eta_b) L_b' + \Psi_b)^{-1} S_b)] \quad (4.35)$$

where:

$Cov(\eta_b)$  :  $k \times k$  covariance matrix of between-level factors

$L_b$  :  $p \times k$  between-level factor loading matrix

$S_b$  :  $p \times p$  sample between-individual covariance matrix

$p$  : Number of variables

$k$  : Number of factors

$G$  : Number of individuals

It should be noted here that the MLE of population between-individual covariance matrix is described by the below equation:

$$S_b^* = c^{-1}(S_b - S_{pw}) \quad (4.36)$$

where:

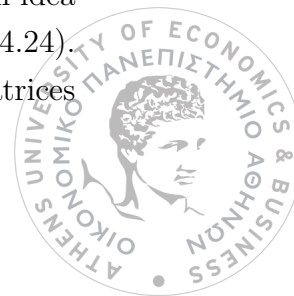
$S_b$  : Biased sample between-individual covariance matrix

$S_{pw}$  : Sample pooled within-individual covariance matrix

$c$  : Average cluster size (see equation 2.7)

$S_b^*$  : Unbiased sample between-individual covariance matrix

Although this estimate is an unbiased estimator of the population between-individual covariance matrix, it produces regularly non-positive definite covariance matrix when conducting factor analysis, so the unbiased estimate is usually used to get a rough idea of the factor structure of the data of the between-individual model (see equation 4.24). If both covariances matrices can produce results, then the structure of both matrices



will be similar (B. O. Muthén, 1994).

## 4.7 Rotation

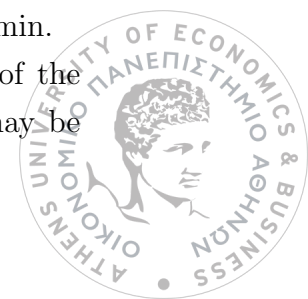
After determining the number of factors in EFA, it is important to use a rotation method that will maximize the interpretability of the model, a property that is referred as simple structure (Thurstone, 1947). Generally, for any factor model there is an infinite number of equivalent solutions, each of one including a different factor loading matrix. So the purpose is to find a matrix that will give the most interpretable results. The criteria for defining this idea was introduced by Thurstone (1974) as follows:

1. Each row of the factor loadings matrix should include at least one zero
2. Each column should contain at least  $k$  zero with  $k \neq 0$
3. Every pair of columns should have several rows with zeros in one column and no zero in the other
4. If  $k \geq 4$ , every pair of columns should have several rows with zeros in both columns
5. Every pair of columns should have few rows with non-zero loadings in both columns

Practically, these rules indicate that in order to attain a simple structure for the factor loading matrix, each factor should be highly correlated with a subset of items and each item should be highly correlated with only one factor (i.e., loading  $> 0.45$ ) (McDonald & Goldstein, 1989).

When someone is interested in finding such a simple structure, there is an important consideration regarding the type of rotation that should be used. Such a consideration is informed based on whether the factor should be assumed to be correlated or not. The first type is called orthogonal rotation and the second type oblique rotation (or non-orthogonal). Orthogonal rotation constraints the factors to be uncorrelated, whereas oblique rotation allows correlation between the factors. The magnitude of such correlation in the rotation is quantified via the cosine of the angle between the rotational axis. So if there is an orthogonal rotation the angle will be  $90^\circ$  as  $\cos(90^\circ)$  is equal to 0, and on the oblique rotation the angle will be either less than  $90^\circ$  or greater than  $90^\circ$ . Typical examples of orthogonal rotation are varimax, quartimax and equimax rotation and some examples of oblique rotation are promax, geomin, quartamin and oblimin.

The differences of the two type of rotations also affects the interpretation of the results. In the case of oblique rotation, the correlation of items and factors may be



inflated by the correlation among factors given what matrix is being used. That is because there are two matrices for factor loadings in such rotation. The first matrix is called structure matrix and the second one pattern matrix. The structure matrix is derived by multiplying the loading with the correlation among factors, so the adjusted loadings include the correlation between items and factors and the correlation among factors. The other matrix includes only the former correlation, and generally this matrix is used in most of the statistical packages (Brown, 2015).

On orthogonal rotation, on the other hand, there is not any inflation of the correlation between the items and factors. However, this may lead to misleading results, as the defined correlation structure of factors may not be representative of the true correlation. For instance, when data are derived from a questionnaire, it is possible the factor structure to include correlated domains (i.e., symptoms in different body areas) that measure a broader concept (i.e., overall symptom severity). So using oblique rotation can give more realistic results in such cases. That is why in the simulation study of this thesis oblique rotation was chosen and more specifically oblimin rotation which tries to achieve a simple structure by minimizing the cross-product of loadings (Tabachnick, Fidell, & Ullman, 2013)

The general idea for rotating the factor model is to minimize a function  $g(V)$  which measures the complexity of the factor loading matrix. Let  $V$  be the factor loading matrix after the rotation. Then this matrix is derived based on the post-multiplication of the initial factor loading  $p \times k$  matrix and  $T$   $k \times k$  matrix. Let also  $F'$  be the correlation among factors after the rotation.

For orthogonal rotation where there is not any constraint on the correlation among factors there are only  $m$  constraints, described in the below equation:

$$Diag(F') = diag(T^{-1}T^{-1}') = I \quad (4.37)$$

where:

$I$  :  $p \times k$  identity matrix

For the case of the oblimin rotation the complexity function is described as follows based Herman (1976) formula:

$$g(V) = \sum_{p < k=1}^K (n \sum_{p=1}^P u_{pk}^2 u_{pk'}^2 - \gamma \sum_{p=1}^P u_{pk}^2 \sum_{p=1}^P u_{pk'}^2) \quad (4.38)$$



where:

$u_{pk}$  : loadings from the  $V$  matrix

$P$  : Number of variables

$K$  : Number of factors

$\gamma$  : constant value which controls the magnitude of the correlation among factors

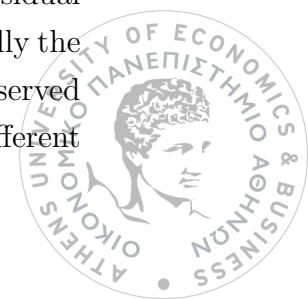
If  $\gamma$  is close to 0 then the factor are highly correlated, which is generally not preferred as values higher than 0.80 or 0.85 indicate poor discriminant validity (Brown, 2015). A more preferred value, based on empirical evidence, is  $\gamma$  equal to 0.5 (Carroll, 1953).

## 4.8 Goodness of fit measures

### 4.8.1 Introduction

After exploring the latent structure of the data with the help of factor analysis, the final goal is to assess the fit of the selected model and more specifically to assess how well the model reproduces the observed covariance matrix. A common goodness of fit statistics that is used for evaluating the fit of the model, is Chi-squared statistic which is often also called either ‘badness of fit’ (Kline, 2015) or ‘lack of fit’ (Mulaik et al., 1989). This measure quantifies the distance between the observed and model based covariance matrix. Such a measure assumes multivariate normality and a large sample size, which often do not hold under real-world datasets (Schermelleh-Engel et al., 2003). In the first instance, where the multivariate normality assumption is not met, the Chi-squared statistic will tend to reject a model even when is correctly specified (Mulaik et al., 1989). In the second instance, the results could be significantly affected for either small or large sample sizes. When the sample size is large, the Chi-squared statistic will usually reject the null hypothesis that the observed and model based covariance matrices are close (Bentler & Bonett, 1980). However, when sample size is small, a bad, and a good fitted model are not always indistinguishable based on that fit statistic (Kenny & McCoach, 2003).

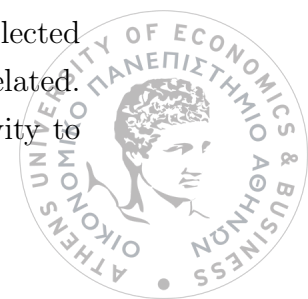
That is why additional goodness of fit measures were proposed, that provide a more holistic view of the assessment of model fit. The first category refers to goodness of fit measures of the overall fit of the model, which include Root Mean Square Residual (RMR) and SRMR, which are a function of the fitted residuals of the model. Usually the most widely preferred measure is SRMR as RMR is affected by the scale of the observed variables. So if for example there is a questionnaire that consists of items with different



scale range, the results are difficult to be interpreted (Kline, 2015). Although SRMR carries this advantage, it still shares some disadvantages of RMR. More specifically, when the number of parameters is large or the sample size is large, the magnitude of SRMR and RMR can be small (Hooper, Coughlan, & Mullen, 2008). This practically means that if there is a model that has numerous parameters or great sample size, SRMR may potentially suggest a good fit due to the overparameterization or the large sample size rather than the actual fit of the model. It also does not take into account the sign of the residuals which could inform, whether the model underestimates or overestimates the sample covariance.

Another measure that could be employed is root mean square error of approximation (RMSEA) which belongs to the broader category of absolute fit indices along with Chi-squared, SRMR and RMR. RSMEA (Steiger, 1980) is generally considered as one of the most well informative goodness of fit measures (Diamantopoulos & Sigauw, 2000). That is because it is affected by the number of parameters of the model, which means that it penalizes models that are less parsimonious. Such a measure tries to measure to what extent the model based on the optimally chosen estimated parameters can reproduce the population covariance (Byrne, 2013), which is why it is called error of approximation as it tries to approximate the level of model misfit relative to the population covariance matrix. It is based on the idea that if the model can reproduce the population covariance matrix adequately then the fitting function that is employed, based on the selected estimation method, will be minimized. Such a measure is lowly dependent on sample size, and it is more supportive of parsimonious models (Browne & Cudeck, 1992). An additional advantage is that it can provide a confidence interval for the population RSMEA (MacCallum, Browne, & Sugawara, 1996), which allows the implementation of a hypothesis testing to assess whether the fit of the model is adequate or not.

Although this measure could be categorized under the same umbrella with SRMR, RMR and Chi-squared, some researchers considered it under the category of the parsimony correction indices (Brown, 2015) along with other measures such as CFI and TLI. That is because all these measures penalize models as the number of parameters increase. However, CFI and TLI have one fundamental difference with RMSEA as they are descriptive measures based on model comparisons. This means that they compare the selected models versus a baseline model, which is usually assumed to have error variances fixed to zero and factor loadings fixed to one. This model is called the independence model, and the only parameters to be estimated are the estimated variances of the variables. So practically, they measure how much better the selected model fits in comparison with a model that assumes that the data are uncorrelated. Both CFI and TLI carry the same advantage as RMSEA regarding the sensitivity to

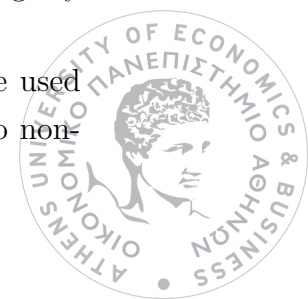


sample size (Bentler, 1990). CFI in particular is one of the most widely used goodness of fit measure in applied research, as is it considered as one of the fit indices that is less affected by sample size (Fan, Thompson, & Wang, 1999). Although TLI carries the same advantages with CFI, under a small sample size, it can indicate a bad fit despite other indices suggesting a good fit, (Bentler, 1990; Kline, 2015) even in cases when the fit of the model is actually good. This measure is also a non-normed index, which means that sometimes the values can get above 1, and this can cause challenges in the interpretation of the results (Byrne, 2013).

Consequently, each goodness of fit measure carry both strengths and weaknesses, and that is why a combination of these measures along with a critical review is required to assess the fit of a factor analysis model. This is also supported by multiple researchers, who strongly suggest to use various goodness of fit measures instead of just one (Mueller, 1999; Crowley & Fan, 1997; Hu & Bentler, 1999; Bollen & Long, 1993). Although there are many suggestions on which goodness of fit measures should be used (Hu & Bentler, 1999; Kline, 2015; Boomsma, 2000), some commonly used are SRMR, TLI, CFI and RMSEA. Hu and Bentler (1999) suggested a 2-index presentation strategy, which always includes the 4 measures mentioned above. They also proposed thresholds for defining an acceptable, bad or good fitting model, including SRMR, TLI, CFI and RMSEA by examining their rejection rates on correctly and misspecified models based on a simulation study. This study was highly influential in Structural Equation Modelling, as many practices still uses some conventional cut-offs.

However, these cut-offs values were build based on continuous data and MLE theory. This means that the conventional cut-off values may not apply for ordinal data. More specifically, there has been a discussion regarding their appropriateness in categorical data in the literature (Shi, Maydeu-Olivares, & Rosseel, 2020; McNeish et al., 2021; Marsh, Hau, & Wen, 2004; Xia & Yang, 2019; Shi & Maydeu-Olivares, 2020). More specifically it was found that when using CFI, RMSEA and TLI in ordinal data under the DWLS estimation, they will suggest a better fit in comparison with MLE even in misspecified models whereas SRMR it was found to be more robust to the estimation method used, especially under large sample size (Shi & Maydeu-Olivares, 2020). Although there has been a discussion and consideration regarding the fact that thresholds could not be universal for both ordinal and continuous data, there have not been any alternative proposals, that is why usually the same cut-off values are used for both type of data. For this thesis in particular the cut-off values for RMSEA, SRMR, TLI and CFI will be based on (Schermelleh-Engel et al., 2003) recent suggestions with slightly small differences with (Hu & Bentler, 1999) recommendations.

For the cases where DWLS estimation method is used, scaled values will be used for CFI, TLI and RMSEA, which utilize Chi-squared statistics that is robust to non-



normality with adjusted mean and variance (Asparouhov & Muthén, 2010). Although there has not been any study justifying the use of the robust chi-squared for calculating the goodness of fit measures, it is widely used by researchers (Savalei, 2014). So unscaled Chi-squared, RMSEA, TLI and CFI will be defined based on MLE method and then scaled Chi-squared, RMSEA, TLI and CFI for DWLS estimation method will be also presented. Finally, SRMR will be defined the same way for both estimation methods, as it is not affected by the Chi-squared statistics.

#### 4.8.2 Unscaled fit statistics for Chi-squared, RMSEA, CFI and TLI

For MLE method the following unscaled statistics are calculated for Chi-squared, RMSEA, CFI and TLI:

$$X_{unscaled}^2(df_H) = (N - 1)F_{ML}[S, \Sigma(\bar{\theta})] \quad (4.39)$$

where:

$df_H$  : Difference between the redundant elements in the covariance matrix  $S$  and the total number of parameters to be estimated for the hypothesized model

$N$  : sample size

$S$  : Observed covariance matrix

$\Sigma(\bar{\theta})$  : Estimated covariance matrix based on the model

$F_{ML}$  : Fitting function based on MLE estimation method

$N$  : Total number of observations

$$RMSEA_{unscaled} = \sqrt{\max\left(\frac{F_{MLE}[S, \Sigma(\bar{\theta})]}{df_H} - \frac{1}{N}, 0\right)} \quad (4.40)$$



where:

$df_H$  : Difference between the redundant elements in the covariance matrix  $S$  and the total number of parameters to be estimated for the hypothesized model

$N$  : sample size

$S$  : Observed covariance matrix

$\Sigma(\bar{\theta})$  : Estimated covariance matrix based on the model

$F_{MLE}$  : Fitting function based on MLE method

$N$  : Total number of observations

Based on the proposed cut-off values (Schermelleh-Engel et al., 2003):

- If  $RMSEA > 0.08$  the fit of the model is bad
- If  $0.05 < RMSEA \leq 0.08$  the fit of the model is acceptable
- If  $0.00 < RMSEA \leq 0.05$  the fit of the model is good

$$CFI_{unscaled} = 1 - \frac{\max(X_{unscaled_H}^2 - df_H, 0)}{\max(X_{unscaled_H}^2 - df_H, X_{unscaled_B}^2 - df_B, 0)} \quad (4.41)$$

where:

$X_{unscaled_H}$  : Unscaled Chi-squared for the hypothesized model

$X_{unscaled_B}$  : Unscaled Chi-squared for the baseline model

(independence model as was explained previously)

$df_H$  : Difference between the number of non-redundant element in  $S$  and the total number of parameters to be estimated based on the hypothesized model

$df_B$  : Difference between the number of non-redundant element in  $S$  and the total number of parameters to be estimated based on the baseline model

$F_{MLE}$  : Fitting function based on MLE method

$N$  : Total number of observations

Based on the proposed cut-off values (Schermelleh-Engel et al., 2003):

- If  $CFI < 0.95$  the fit of the model is bad



- If  $0.95 \leq CFI < 0.97$  the fit of the model is acceptable
- If  $0.97 \leq CFI < 1.00$  the fit of the model is good

$$TLI_{unscaled} = 1 - \frac{\frac{X_{unscaled_B}^2}{df_B} - \frac{X_{unscaled_H}^2}{df_H}}{\frac{X_{unscaled_B}^2}{df_B} - 1} \quad (4.42)$$

where  $X_{unscaled_H}$ ,  $X_{unscaled_B}$ ,  $df_H$ ,  $df_B$  are predefined below.

Based on the proposed cut-off values (Schermelleh-Engel et al., 2003):

- If  $TLI < 0.95$  the fit of the model is bad
- If  $0.95 \leq TLI < 0.97$  the fit of the model is acceptable
- If  $0.97 \leq TLI < 1.00$  the fit of the model is good

### 4.8.3 Scaled fit statistics for Chi-squared, RMSEA, CFI and TLI

For DWLS estimation method the following scaled statistics are calculated for Chi-squared, RMSEA, CFI and TLI:

$$X_{scaled}^2(df) = \frac{X_{unscaled}^2}{a} + b \quad (4.43)$$

where:

$df_H$  : Difference between the redundant elements in the covariance matrix  $S$  and the total number of parameters to be estimated for the hypothesized model

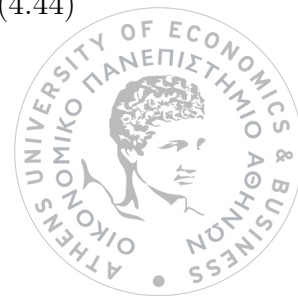
$X_{unscaled}$  : Unscaled Chi-squared statistic as previously defined

$\alpha$  : Scaling parameter

$\beta$  : Shifting parameter

$\alpha$  and  $\beta$  are derived in such a way so that  $E(X_{scaled}^2) = df_H$  and  $Var(X_{scaled}^2) = 2df_H$  (for more details, see (Asparouhov & Muthén, 2010)).

$$RMSEA_{scaled} = \sqrt{\max\left(\frac{F_{WLS}[S, \Sigma(\bar{\theta})]}{df_H} - \frac{1}{N}, 0\right)} \quad (4.44)$$



where:

$df_H$  : Degrees of freedom for the hypothesized model

$N$  : Number of observations

$F_{DWLS}$  : Fitting function for DWLS

$\alpha$  and  $\beta$  are derived in such a way so that  $E(X_{scaled}^2)=df_H$  and  $Var(X_{scaled}^2)=2df_H$  (for more details, see (Asparouhov & Muthén, 2010)).

$$CFI_{scaled} = 1 - \frac{\max(X_{scaled_H}^2 - df_H, 0)}{\max(X_{scaled_H}^2 - df_H, X_{scaled_B}^2 - df_B, 0)} \quad (4.45)$$

where:

$X_{scaled_H}$  : Scaled Chi-square statistics for the hypothesized model

$X_{scaled_B}$  : Scaled Chi-square statistics for the baseline model

$df_H$  : Degrees of freedom for the hypothesized model

$df_B$  : Degrees of freedom for the baseline model

$F_{DWLS}$  : Fitting function for DWLS

$$TLI_{scaled} = 1 - \frac{\frac{X_{scaled_B}^2}{df_B} - \frac{X_{scaled_H}^2}{df_H}}{\frac{X_{scaled_B}^2}{df_B} - 1} \quad (4.46)$$

where:

$X_{scaled_H}$  : Scaled Chi-square statistics for the hypothesized model

$X_{scaled_B}$  : Scaled Chi-square statistics for the baseline model

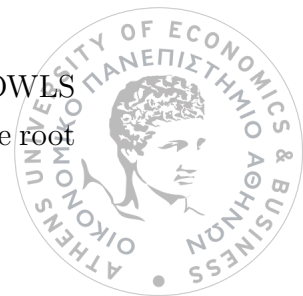
$df_H$  : Degrees of freedom for the hypothesized model

$df_B$  : Degrees of freedom for the baseline model

$F_{DWLS}$  : Fitting function for DWLS

#### 4.8.4 SRMR

Given that SRMR is not affected by the Chi-squared function, for both ML and DWLS estimation methods the formula will be the same. SRMR is described as the square root



of the average discrepancy between the observed correlation matrix and the estimated correlation matrix based on the model. It can be derived as follows:

$$SRMR = \sqrt{\frac{(r_{ij} - \hat{r}_{ij})^2}{d}} \quad (4.47)$$

where:

$r_{ij}$  : Estimated correlation between the  $i$  and  $j$  variable

$\bar{r}_{ij}$  : Observed correlation between the  $i$  and  $j$  variable

$d$  : Number of pairs between the variables

Based on the proposed cut-off values (Schermelleh-Engel et al., 2003):

- If  $0.1 < SRMR$  the fit of the model is bad
- If  $0.05 < SRMR \leq 0.1$  the fit of the model is acceptable
- $SRMR \leq 0.05$  the fit of the model is good

The above goodness of fit measures serve as useful tools to assess the fit of the EFA model. Given the unknown factor structure of the data, they enable the investigation of how well the structure that was found is representative of the observed data. In other terms, they can define how well the selected model can reproduce the covariance matrix. Such a purpose is perfectly aligned with another model that will be investigated under this thesis: the CFA model.



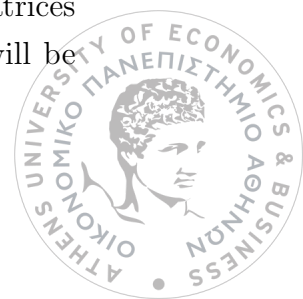
# Chapter 5

## Confirmatory factor analysis

### 5.1 Introduction

As with uses of EFA, CFA is used to identify factors that contribute to the variation and covariation of the observed variables. These two frameworks are built based on the same model, with slight differences. That is why the terminology used in the previous Chapter will be the same for this Chapter as well. The differences between these two models is that the EFA model does not have any restriction on the parameters. That is why this model is used when there is no a priori knowledge of the subgroup of items that load to each factor. So the loadings of each item in EFA will be calculated for every factor, even though in reality the item might load to only one factor. When knowledge regarding model properties is already available known via prior evaluation/hypothesis, the CFA model is the right choice. Based on this model, someone can prespecify various aspects of the model, such as the number of factors and the association strength between items and factors. So if for example someone knows that a specific item loads to factor 1 only then the corresponding factor loadings on the rest of the factors can be set to 0. This is an important difference between EFA and CFA models, as the one is used for exploratory purposes whereas the other one is used for the confirmation of the factor structure that is hypothesized based on previous empirical evidence. Consequently, such a model is usually used after later stages of scale development or structural validity which is established based on EFA models or based on theoretical background (Brown, 2015).

In this Chapter, a brief discussion regarding the identification issues of such models will be discussed. Then the CFA model for the single selected day, item weekly average and multilevel approaches will be described along with their input correlation matrices and the corresponding estimation methods. Finally, goodness of fit measures will be defined as well.



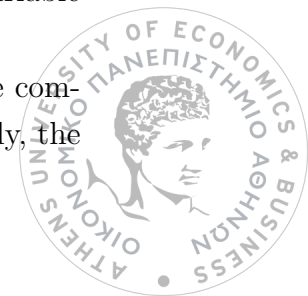
## 5.2 Identification of the confirmatory factor analysis model

An important rule that should be held for a CFA model in order the corresponding parameters to be estimated is that it should be identifiable. This means that based on the available information provided by covariance/correlation matrix, one should obtain unique values for each of the unknown parameters (i.e. factor loadings, error variances). Identification is usually quantified via the difference of the number of constant values that are needed for the estimation of the model and the freely estimated parameters of the model.

More specifically, when the number of freely estimated parameters is equal to the number of values of the input covariance then the model is called just identified, whereas when the number of freely estimated parameters exceed the number of values of the input covariance matrix then the model is under-identified and the estimated parameters can not be estimated. In the case where there are less unknown parameters to be estimated than the available values of the input covariance matrix, then the model is over-identified. If the model is either just identified or over-identified, the model parameters could be estimated. However, over-identified models are needed in order to obtain goodness of fit measures, which are necessary for the evaluation of model fit. That is because when a model is over-identified, then there are an infinite number of possible equivalent solutions, so there is a possibility that the proposed solution may be wrong and this ensures that there will be an error in the model. If for every piece of available information there was a parameter to be estimated, then the fit of the model would be perfect, as every parameter would be perfectly estimated. So there would not be any point for the evaluation of such a model.

With this central issue of the identification of the model, there is also an important consideration regarding the identification of the latent variables which are not directly observed. Due to the fact that latent variables do not have a known scale, it is necessary to fix their variance into some constant values, directly or indirectly. In the first case, which is the direct way, one can fix the factor variances equal to 1 and proceed to the estimation of the model parameters. The second way that indirectly fixes the variance of the latent variable, is to fix the first factor loading of each factor equal to 1. If the corresponding loadings are equal to 1, then the factor is perfectly associated with the corresponding item, which means that any change in the observed variable will result in an equivalent change to the factor. So indirectly, the variance of the observed variable is passed to the variance of the latent variable.

Consequently, one of the two restrictions should be imposed regardless of the complexity of the model, as it is necessary for the model to be identified. Additionally, the



number of values of the input covariance matrix must be equal or exceed the number of parameters, as was previously mentioned. In the special case of a one-factor model, it is required to have at least 3 available indicators (Brown, 2015). If the number of indicators is 3, and it is assumed that there is not any correlation between the error, the one-factor model is just-identified, so no goodness of fit measures could be employed as was previously explained. In the case where there are more than 3 indicators, the model would be over-identified and model fit evaluation could be utilized. In case there are at least 2 factors in the model with at least two indicators within each factor, the model is over-identified, with the presumption that each factor is correlated with at least one of the other factors and the errors of the model are uncorrelated. However, empirically such a rule could still result in under-identification, that is why 3 items per factors are required.

## 5.3 Confirmatory ordinal factor analysis for single selected day approach

### 5.3.1 Input correlation matrix: polychoric correlation

As in the case of exploratory ordinal factor analysis, the input correlation matrix will be defined based on the polychoric correlation, which is a correlation measure applicable to ordinal data. For more details, see Section 4.4.1

### 5.3.2 Model

Let  $y_{ti}$  be a  $p$ -dimensional vector, including the observed variables at time  $t$  for individual  $i$ . Let's also assume that the first  $a_1$  observed variables load on factor 1, the next  $a_2$  variables load on factor 2, ..., the last  $a_k$  variables load on factor  $k$ .

The form of the model is described as follows:

$$y_{ti}^* = L^* \eta_i^* + \epsilon_{ti}^* \quad (5.1)$$



where:

$L^*$  :  $pxk$  factor loading matrix where the first  $a_1$  rows in the first column are non zero values and the rest of the rows include only zeros , the second column includes non zero values only on  $a_1+1$  until  $a_2$  row and so on.

$p$  : Number of variables

$k$  : Number of factors

$y_{ti}^*, \eta_i^*, \epsilon_{ti}^*$  : They are defined the same way as defined in EFA (see Section 4.4.2)

As previously discussed in the EFA Chapter, the model could be also described through the covariance matrix as follows:

$$\Sigma^* = L^*Cov(\eta^*)L^* + \Psi^* \quad (5.2)$$

where:

$L^*$  :  $pxk$  factor loading matrix where the first  $a_1$  rows in the first column are non zero values and the rest of the rows include only zeros , the second column includes non zero values only on  $a_1+1$  until  $a_2$  row and so on

$p$  : Number of variables

$k$  : Number of factors

$\Sigma^*, L^*Cov(\eta^*)L^*, \Psi^*$  : They are defined the same way as defined in EFA (see Section 4.4.2)

In order for the model to follow the identification rules (Chen et al., 2001), the variance of the factors should be equal to 1. Additionally, the variance of measurement error is not identified. That is why in ordinal factor analysis the model can be identified by setting  $\Psi^*$  as described below:

$$\Psi^* = I - \text{diag}(L^{*'}Cov(\eta)L^*) \quad (5.3)$$

Note that this model assumes that there is correlation among factors, so this model is non-orthogonal, whereas orthogonal models assumes that there is no correlation among factors.



### 5.3.3 Distributional assumptions

The assumptions of the CFA model are the same as in the EFA (see Section 4.4.3).

### 5.3.4 Additional assumptions

The assumptions of the CFA model are the same as in the EFA (see Section 4.4.4).

### 5.3.5 Estimation method

Given that the data are ordinal in the single selected day approach, the same estimation method is used as in in the case of EFA which is the DWLS estimation method (see Section 4.4.5).

## 5.4 Confirmatory factor analysis for item average approach

### 5.4.1 Input correlation matrix: Pearson correlation matrix

The second approach refers to averaging the item across time for each individual, and it is referred to as the item average approach. Based on the simulation study, the data are on an ordinal scale, so averaging the items across time will result in transforming the data from an ordinal to continuous scale, so Pearson correlation will be used as in the case of EFA. For more details, see Section 4.5.1.

### 5.4.2 Model

Given that the average weekly item approach transforms the ordered data to continuous, common CFA could be employed by using as input correlation matrix Pearson correlation matrix. Let  $y_{ti}$  be a  $p$ -dimensional vector, including the observed variables at time  $t$  for individual  $i$ . Let's also assume that the first  $a_1$  observed variables load on factor 1, the next  $a_2$  variables load on factor 2, ..., the last  $a_k$  variables load on factor  $k$ .

The form of the model is described as follows:

$$\bar{y}_{.i} = L\eta_i + \epsilon_i \quad (5.4)$$



where:

$L$  :  $pxk$  factor loading matrix where the first  $a_1$  rows in the first column are non zero values and the rest of the rows include only zeros , the second column includes non zero values only on  $a_1+1$  until  $a_2$  row and so on.

$p$  : Number of variables

$k$  : Number of factors

$\bar{y}_i, \eta_i, \epsilon_i$  : They are defined the same way as defined in EFA (see Section 4.5.2)

Although this form of the model is adequate to describe the CFA model, an alternative way is to describe it through the covariance matrix as in the EFA framework.

$$\Sigma = LCov(\eta)L' + \Psi \quad (5.5)$$

where:

$L$  :  $pxk$  factor loading matrix where the first  $a_1$  rows in the first column are non zero values and the rest of the rows includes only zeros , the second column includes non zero values only on  $a_1+1$  until  $a_2$  row and so on.

$p$  : Number of variables

$k$  : Number of factors

$\Sigma, LCov(\eta)L', \Psi$  : They are defined the same way as defined in EFA (see Section 4.5.2)

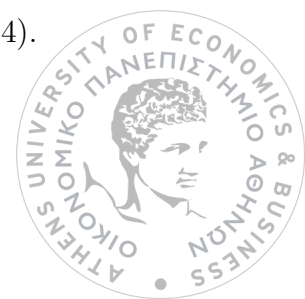
Note that this model assumes that there is correlation among factors, so this model is non-orthogonal, whereas orthogonal models assumes that there is no correlation among factors.

### 5.4.3 Distributional assumptions

The assumptions of the CFA model are the same as in the EFA (see Section 4.5.3).

### 5.4.4 Additional assumptions

The assumptions of the CFA model are the same as in the EFA (see Section 4.5.4).



### 5.4.5 Estimation method

Given that the data are continuous in the weekly item average approach, the same estimation method is used as in the case of EFA which is the MLE method (see Section 4.5.5).

## 5.5 Confirmatory multilevel factor analysis

The multilevel approach is referring to the decomposition of the total covariance matrix to within- and between-individual covariance matrix by estimating within- and between-individual parameters simultaneously through a CFA model. So in the confirmatory framework where the within- and between-individual analysis are conducted simultaneously, the input covariance matrix will be both within- and between-individual covariance matrix.

### 5.5.1 Input covariance matrix: within and between-individual covariance matrix

The input covariance matrix for the within-individual part of the multilevel model will be:

$$S_{pooled} = \frac{\sum_{i=1}^n \sum_{t=1}^{t_i} (y_{ti} - \bar{y}_{.i})(y_{ti} - \bar{y}_{.i})'}{N - G} \quad (5.6)$$

where  $n$ ,  $N$ ,  $G$ ,  $t_i$ ,  $y_{ti}$ , and  $\bar{y}_{.i}$  are defined the same way as in the EFA (see Section 4.6.1).

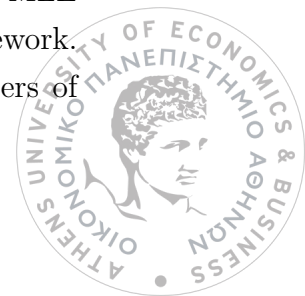
The input covariance matrix for the between-individual part of the multilevel model will be:

$$S_{between} = \frac{\sum_{i=1}^n (\bar{y}_{..} - \bar{y}_{.i})(\bar{y}_{..} - \bar{y}_{.i})'}{n - 1} \quad (5.7)$$

$\bar{y}_{..}$  and  $\bar{y}_{.i}$  are defined the same way as in the EFA (see Section 4.6.2).

### 5.5.2 Model

In this approach, all the data are utilized, and multilevel factor analysis is employed by using as input covariance matrices both within- and between-individual covariance matrices. The theory behind the decomposition of the data was mainly built on continuous data, so although the data are on an ordinal scale, multilevel factor analysis under MLE theory will be employed. This method was also implemented for the EFA framework. However, for the CFA framework, the within- and between-individual parameters of the model are estimated simultaneously.



Let  $y_{ti}$  be a  $p$ -dimensional vector, including the observed variables at time  $t$  for the individual  $i$ . Let's also assume that the first  $a_{w_1}$  observed variables load on within-level factor 1, the next  $a_{w_2}$  variables load on within-level factor 2, . . . , the last  $a_{w_k}$  variables load on within-level factor  $k$ , the first  $a_{b_1}$  observed variables load on between-level factor 1, the next  $a_{b_2}$  variables load on between-level factor 2, . . . , the last  $a_{b_k}$  load on between-level factor  $k$ .

The form of the model is described as follows, for day  $t$  :

$$y_{ti} = L_b \eta_{b_i} + L_w \eta_{w_{ti}} + \epsilon_{w_{ti}} + \epsilon_{b_i} \quad (5.8)$$

where:

$L_b$  :  $p \times k$  between-level factor loading matrix where the first  $a_{b_1}$  rows in the first column are non zero values and the rest of the rows include only zeros, the second column includes non zero values only on  $a_{b_1}+1$  until  $a_{b_2}$  row and the rest of the rows include only zeros, and so on

$L_w$  :  $p \times k$  within-level factor loading matrix where the first  $a_{w_1}$  rows in the first column are non zero values and the rest of the rows include only zeros, the second column includes non zero values only on  $a_{w_1}+1$  until  $a_{w_2}$  row and the rest of the rows include only zeros, and so on

$p$  : Number of variables

$k$  : Number of factors

$\eta_{w_{ti}}, \eta_{b_i}, \epsilon_{w_{ti}}$  : They are defined the same way as defined in EFA (see Section 4.6.3)

Note that for simplicity is it assumed the number of factors is the same across the two levels.

Alternatively, the model could be described based on the within- and between-individual covariance matrix as follows:

$$\Sigma_w = L_w Cov(\eta_w) L_w' + L_b Cov(\eta_b) L_b' + \Psi_b + \Psi_w \quad (5.9)$$



where:

$L_b$  :  $pxk$  between-level factor loading matrix where the first  $a_{b_1}$  rows in the first column are non zero values and the rest of the rows include only zeros, the second column includes non zero values only on  $a_{b_1}+1$  until  $a_{b_2}$  row and the rest of the rows include only zeros, and so on

$L_w$  :  $pxk$  within-level factor loading matrix where the first  $a_{w_1}$  rows in the first column are non zero values and the rest of the rows include only zeros, the second column includes non zero values only on  $a_{w_1}+1$  until  $a_{w_2}$  row and the rest of the rows include only zeros, and so on

$p$  : Number of variables

$k$  : Number of factors

$\Sigma_w, L_w Cov(\eta_w) L_w$  : They are defined the same way as defined in EFA (see Section 4.6.3)

$\Sigma_b, L_b Cov(\eta_b) L_b$  : They are defined the same way as defined in EFA (see Section 4.6.3)

For this model, the first loading within each factor is fixed to one across both levels of the model. Additionally, there are constraints regarding equality of factor loadings, variance, and covariances of the observed variables and the factors for the within-individual model which is included at both levels of the model (i.e., the within-individual model parameters are constrained to be equal across both within- and between-individual level).

### 5.5.3 Distributional assumptions

The assumptions of the CFA model are the same as in the EFA (see Section 4.6.4).

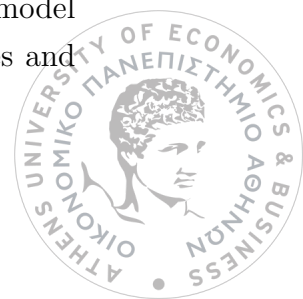
### 5.5.4 Additional assumptions

The assumptions of the CFA model are the same as in the EFA (see Section 4.6.5).

### 5.5.5 Estimation method

For implementing the multilevel CFA, MLE is utilized.

The estimated parameters for the confirmatory multilevel factor analysis model are within- and between-level loadings, within- and between-level error variances and



within and between-level factor variances and covariances given that the model is non-orthogonal.

The equation 5.10 describes the probability of observing the sample within- and between-individual covariance matrices simultaneously given the estimated parameters in log scale for confirmatory multilevel factor analysis model with correlated factors. This likelihood is maximized with respect to the estimated within- and between-level loadings, within- and between-level error variances and within- and between-level factor variances and covariances.

$$\begin{aligned}
 F_{MLE} = L(y_{ti}, L_w, L_b, \Psi_w, \Psi_b, \eta_w, \eta_b) = & G(\log(|L_w Cov(\eta_w)L'_w + \Psi_w + c(L_b Cov(\eta_b)L'_b + \Psi_b)|) \\
 & + tr((L_w Cov(\eta_w)L'_w + \Psi_w + c(L_b Cov(\eta_b)L'_b + \Psi_b)^{-1}S_b) \\
 & - \log |S_b| - p + (N - G) * (\log |\Sigma_w| + tr((L_w Cov(\eta_w)L'_w \\
 & + \Psi_w)^{-1}S_w) - \log |S_w| - p
 \end{aligned}
 \tag{5.10}$$

where:

$Cov(\eta_w)$  : Covariance matrix between within-level factors

$Cov(\eta_b)$  : Covariance matrix among between-level factors

$L_b$  :  $p \times k$  between-level factor loading matrix where the first  $a_{b_1}$  rows in the first column are non zero values and the rest of the rows include only zeros, the second column includes non zero values only on  $a_{b_1}+1$  until  $a_{b_2}$  row and so on

$L_w$  :  $p \times k$  within-level factor loading matrix where the first  $a_{w_1}$  rows in the first column are non zero values and the rest of the rows include only zeros, the second column includes non zero values only on  $a_{w_1}+1$  until  $a_{w_2}$  row and so on

$S_w$  : Covariance matrix among within-level factors

$S_b$  : Covariance matrix among between-level factors

$N$  : Total number of observations

$G$  : Number of individuals

$c$  : Average cluster size (see equation 2.7)

$\Psi_w$  : Within-level specificity

$\Psi_B$  : Between-level specificity



## 5.6 Goodness of fit measures

Given that the same estimation methods is used as in the exploratory framework, the same goodness of fit measures will be used for the evaluation of model fit. For more details see Chapter 4.8.

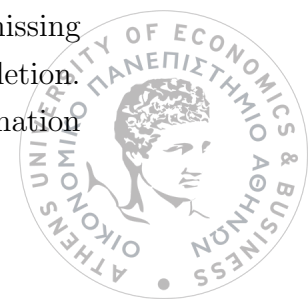
## 5.7 Heywood cases

An issue that often arises in factor analysis, is when the parameter estimates produce out of range values. This problem is known as Heywood case (Harman & Fukuda, 1966). The two most common examples are when an error variance estimate is negative, or when the standardized loading is greater than 1. Such issues can be really crucial as they might lead to improper solutions (Brown, 2015).

An important condition that is required to obtain proper solutions is the observed covariance or correlation matrix of the data and the corresponding model-based estimated matrix should be positive definite. When this is not the case, the factor modelling could result in Heywood cases. The cause of this is that when some observed variables are perfectly related, the observed covariance or correlation matrix will produce eigenvalues very close to zero, so the matrix will be non-invertible. As a result of this, any statistic that necessitate the estimation of the inverse of that matrix will not be calculable. In such cases, a simple method is to remove one of the variables that causes the problem of multicollinearity.

Another reason the matrix may be non-positive definite is that the sample size is low. Heywood cases are generally sample size dependent and the likelihood of its occurrence is very high when the sample size is 100 or less and very low when the sample size is 500 or more (Bartholomew, Knott, & Moustaki, 2011). The small sample size can result in a singular matrix due to sampling error and as a result this could cause improper solutions as was previously flagged. In such cases where the error variance might be is negative, one could fix the variance to zero or to a value close to zero (Brown, 2015). In small sample sizes, outliers can also impact the results of the model, as in such cases the influence of the outlier in the results is highly increased. In addition to this, overparametrized models such as multilevel models within the context of small sample size could also result in Heywood cases.

Another possible reasons of the occurrence of Heywood cases could be due to the non convergence of the model due to the underidentification or the misspecification of the model. Finally, Heywood case also occurs when there is large number of missing data and someone uses a simple missing data handling strategy such as listwise deletion. In such cases it is recommended to use more advanced methods such as full information



MLE.

Heywood cases, which frequently occur in factor analysis, hold significant importance, necessitating a comprehensive understanding of its origin in order to proceed to the proper solution. Although Heywood cases were identified in the models under evaluation for this study, further analysis did not proceed towards their solution.



# Chapter 6

## Results

The focus of this Chapter is to provide a detailed presentation of the results from the simulation study which was described in Chapter 3, using both EFA and CFA frameworks.

### 6.1 EFA results

For the EFA framework, the following results are presented:

- Percentage of correct identification of number of factors for the single selected day, weekly item average and multilevel split approaches.
- Factor loading strength and range
- ICC results
- Estimated, true and observed inter-item correlation
- Overall factor loading bias and MSE
- Goodness of fit measures



### 6.1.1 Percentage of correct identification of number of factors

Figure 6.1 presents the percentage of simulated datasets for which the number of factors (three as per the simulation) were correctly identified with each of the data handling approaches under the different selected scenarios (as described in Table 3.2). For both the weekly item average approach and the between-individual element of the multilevel approach, the number of factors was correctly identified for all three inferential criteria across all iterations of the simulated data under all the selected scenarios, with the percentage remaining constantly equal to 100%. With the single selected day method, the parallel analysis approach functioned well, achieving a high percentage accuracy for scenario 2 across all sample sizes and for scenario 1 across sample sizes greater than 100, whereas the level of correct identification was lower for the standard and empirical Kaiser criteria. This level of incorrect identification was well-reflected across scenario 1 where the true indicator patterns were weaker than scenario 2. Even parallel analysis underperformed when the sample size was 100 for scenario 1. A similar pattern was observed in the within-individual analysis across both scenarios. When comparing Kaiser and empirical Kaiser criterion, the latter performed better under small sample sizes ( $n=100$ ,  $n=150$ ,  $n=200$ ) under both scenarios for the single selected day and the within-individual analysis and as the sample size increased they showed a similar performance.

In aggregate, across all data handling approaches, the parallel analysis demonstrated a consistently high level of accuracy, but the performance was more variable for the Kaiser and empirical Kaiser criteria. It was also evident that under small sample sizes, empirical Kaiser criterion performed better than Kaiser criterion. Additionally, the percentages of incorrect identification of the number of factors was a greater issue for scenario 1, with Kaiser criterion having a quite low percentage of correct identification when sample size was 100.

Note that for all approaches the sample size is equivalently specified with the number of individuals ( $n$ ) except for within-individual analysis which utilizes all the available dataset across a 1-week period ( $7*n$ ).



Figure 6.1: Bar plots for the percentage of iterations (1,000 simulated datasets) in which the factor identification correctly identified the simulated number of factors for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). Single day=Single selected day approach; Within=Within-individual analysis; Between=Between-individual analysis; a=slope parameter from multidimensional graded response model.



### 6.1.2 Factor loading strength and range

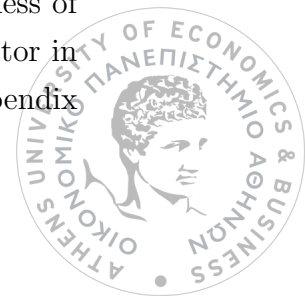
Following implementation of EFA with the simulated datasets, an inspection of the factor loadings showed that the EFA estimated loadings based on the weekly item average approach produced loading estimates that were higher and less variable than the estimates obtained with EFA based on the single day approach and frequently higher than the true parameters, indicating that weekly item average approach is prone to overestimation. For instance, the loadings for the weekly item average method ranged from 0.63-0.98 for the first factor [F1, n=100] for scenario 1 (see Table 6.1). In contrast, the loadings produced with the single day approach had notably wider ranges (i.e., 0.06-0.94 [F1, n=100]). While the ranges were somewhat narrower at higher sample sizes, they remained relatively wider for the single day approach than that based on the weekly item average (i.e., 0.37-0.84 for single day approach [F1, n=350] for scenario 1 and 0.68-0.96 for weekly item average approach [F1, n=350] for scenario 1). The increase of the sample size (i.e., more accurate reflection of the association pattern between items) also led to more accurate results, as the range of the estimated loadings was closer to the range of loading parameters.

The difference between these two approaches was also apparent in the second scenario where the strength of the loadings was higher in comparison with the first scenario (i.e., 0.20-1.01 for single day approach [F1, n=100] and 0.75-1.02 for the weekly item average approach [F1, n=100]). Although a similar pattern appeared, a problem that arose in this scenario was Heywood cases, in which some standardized loadings were higher than 1. This problem possibly occurred as a result of the small number of individuals and the high correlation between the items. However, for larger sample sizes (i.e., n=200 and n=250) this problem did not occur.

The between-individual analysis (i.e., EFA based on the between-individual covariance matrix) produced the same results as the weekly item average approach. However, the within-individual analysis produced weaker loadings than all other approaches and in some occasions it even produced negative values (i.e., -0.01-0.56 [F1, n=100] for scenario 1], 0.29-0.79 [F1, n=100] for scenario 2). As expected, when the sample size was increased the estimated ranges were narrower (i.e., 0.19-0.53 [F1, n=350, scenario 1], 0.27-0.70 [F1, n=350, scenario 2]).

Across the simulations, the within-individual approach produced notably lower estimates and more frequently than the other data handling approaches (for the second and third factor outputs, see Appendix A).

Boxplots across both scenarios were also generated to clarify the distinctiveness of item loading patterns for each approach (an example is provided for the first factor in Figure 6.2, with visualizations for the second and third factor included in the Appendix



B).

Additional analysis was performed to check whether the items loaded on their corresponding factor. In scenario 1 around 450 in 1,000 (45.00%, F1, n=100, scenario 1), 329 in 1,000 (32.90%, F2 n=100, scenario 1), 149 in 1,000 (14.9%, F3, n=100, scenario 1) of the analysis iterations using the single selected day approach had at least one item that did not load on their corresponding factor. As expected the increase of the sample size resulted in lower percentages (i.e., 1.1%, 0.7% 0.8% for F1, F2, F3 respectively, n=350). In scenario 2, given that the true loadings were notably higher in comparison with scenario 1, the percentages were very low even when n was equal to 100 (i.e., 0.9% 1.4% 0.6% for F1, F2, F3 respectively). For the within-individual analysis, at least one of the loadings in 99.80% of iterations for scenario 1 with n=100 did load to its corresponding factor. This pattern also occurred across all the selected sample size, with the percentage being at least 99.60%. For scenario 2 the percentage remained high across the 3 factors but lower in comparison with scenario 1 (i.e., 43.20%, 94.70% and 97.5%, n=100). Similarly, these percentages remained higher even in greater sample sizes (i.e., 93.80%, 97.70% and 24.00%, n=300).

Collectively, weekly item average approach produced notably higher loadings than all the other approaches and was prone to overestimation. The same results were also reflected from the between-individual analysis estimates with the weekly item average approach, as they both study between-individual differences. Scenario 2 also demonstrated that standardized loadings were higher than 1 in some occasions, resulting in the occurrence of Heywood cases. A single selected day produced estimates that were more variable, with a notably wide range of the factor loadings across the 1,000 simulated datasets. As a result, in some datasets some items did not load to their corresponding factor, a problem which occurred more frequently on scenario 1. Within-individual analysis factor loading estimates were also weaker than all the other approaches.



Figure 6.2: Boxplot of the loadings within the first factor across 1,000 simulated datasets for EFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day approach; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.

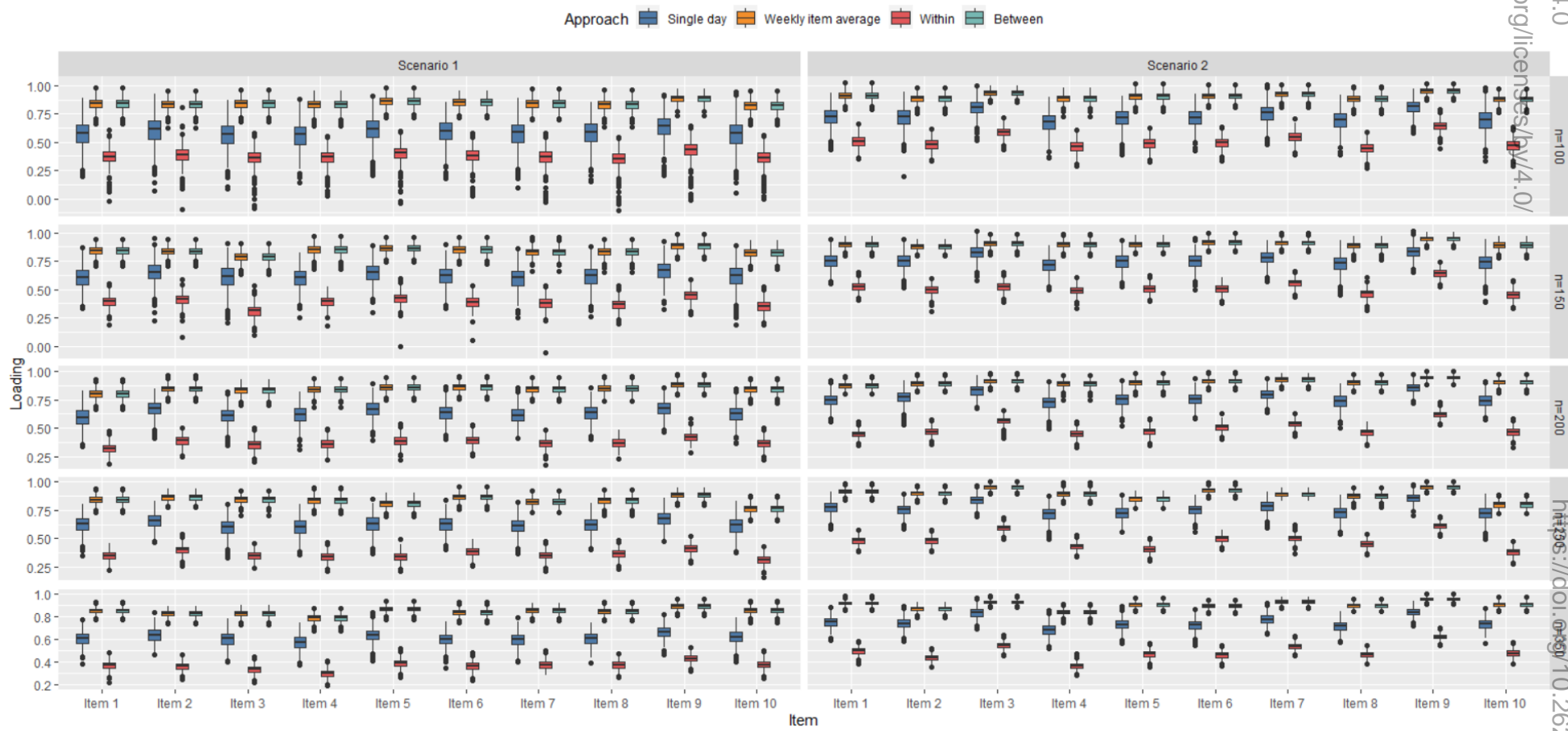


Table 6.1: True loadings parameters, and EFA<sup>a</sup>estimated loadings within factor 1 for the single selected day, weekly item average, within- and between-individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

n <sup>b</sup>	Item	Scenario 1					Scenario 2				
		$L_T^c$	$L_M^d$				$L_T^c$	$L_M^d$			
			SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>		SD <sup>e</sup>	WIA <sup>e</sup>	WA <sup>g</sup>	BA <sup>h</sup>
100	1	0.63	0.20 - 0.89	0.66 - 0.98	-0.02 - 0.61	0.66 - 0.98	0.77	0.44 - 0.94	0.78 - 1.02	0.35 - 0.67	0.78 - 1.02
	2	0.68	0.07 - 0.93	0.63 - 0.95	-0.10 - 0.81	0.63 - 0.95	0.77	0.20 - 0.94	0.75 - 0.98	0.33 - 0.62	0.75 - 0.98
	3	0.62	0.09 - 0.87	0.68 - 0.96	-0.08 - 0.58	0.68 - 0.96	0.85	0.52 - 1.00	0.85 - 1.00	0.43 - 0.72	0.85 - 1.00
	4	0.62	0.14 - 0.88	0.65 - 0.96	0.03 - 0.56	0.65 - 0.96	0.72	0.36 - 0.90	0.73 - 0.98	0.29 - 0.61	0.73 - 0.98
	5	0.67	0.21 - 0.91	0.71 - 0.98	-0.04 - 0.59	0.71 - 0.98	0.75	0.38 - 0.91	0.78 - 1.01	0.32 - 0.63	0.78 - 1.01
	6	0.65	0.18 - 0.85	0.72 - 0.95	0.02 - 0.58	0.72 - 0.95	0.77	0.42 - 0.92	0.81 - 1.00	0.33 - 0.63	0.81 - 1.00
	7	0.63	0.10 - 0.89	0.70 - 0.97	-0.03 - 0.58	0.70 - 0.97	0.80	0.48 - 1.01	0.81 - 1.01	0.39 - 0.71	0.81 - 1.01
	8	0.64	0.15 - 0.85	0.64 - 0.96	-0.10 - 0.54	0.64 - 0.96	0.74	0.39 - 0.92	0.75 - 0.98	0.27 - 0.59	0.75 - 0.98
	9	0.70	0.21 - 0.92	0.73 - 0.97	-0.01 - 0.65	0.73 - 0.97	0.86	0.58 - 0.97	0.87 - 1.02	0.44 - 0.79	0.87 - 1.02
	10	0.64	0.06 - 0.94	0.65 - 0.96	-0.01 - 0.56	0.65 - 0.96	0.75	0.34 - 0.98	0.76 - 0.98	0.29 - 0.64	0.76 - 0.98
150	1	0.63	0.34 - 0.87	0.71 - 0.94	0.19 - 0.55	0.71 - 0.94	0.77	0.55 - 0.94	0.80 - 0.97	0.40 - 0.65	0.80 - 0.97
	2	0.68	0.22 - 0.95	0.70 - 0.94	0.08 - 0.59	0.70 - 0.94	0.77	0.52 - 0.94	0.80 - 0.94	0.31 - 0.60	0.80 - 0.94
	3	0.62	0.21 - 0.90	0.63 - 0.91	0.10 - 0.53	0.63 - 0.91	0.85	0.58 - 1.02	0.81 - 0.99	0.39 - 0.65	0.81 - 0.99
	4	0.62	0.25 - 0.83	0.68 - 0.97	0.18 - 0.52	0.68 - 0.97	0.72	0.50 - 0.88	0.79 - 0.98	0.33 - 0.60	0.79 - 0.98
	5	0.67	0.30 - 0.89	0.74 - 0.96	-0.01 - 0.60	0.74 - 0.96	0.75	0.52 - 0.94	0.81 - 0.98	0.39 - 0.62	0.81 - 0.98
	6	0.65	0.33 - 0.90	0.72 - 0.96	0.05 - 0.53	0.72 - 0.96	0.77	0.50 - 0.95	0.83 - 0.99	0.38 - 0.61	0.83 - 0.99
	7	0.63	0.25 - 0.86	0.66 - 0.96	-0.06 - 0.53	0.66 - 0.96	0.80	0.57 - 0.93	0.83 - 1.00	0.43 - 0.66	0.83 - 1.00
	8	0.64	0.26 - 0.88	0.65 - 0.95	0.20 - 0.53	0.65 - 0.95	0.74	0.45 - 0.93	0.76 - 0.97	0.31 - 0.60	0.76 - 0.97
	9	0.70	0.33 - 0.92	0.76 - 0.99	0.28 - 0.59	0.76 - 0.99	0.86	0.65 - 1.01	0.87 - 1.01	0.52 - 0.75	0.87 - 1.01
	10	0.64	0.19 - 0.89	0.68 - 0.93	0.18 - 0.52	0.68 - 0.93	0.75	0.39 - 0.94	0.78 - 0.97	0.33 - 0.58	0.78 - 0.97
200	1	0.63	0.34 - 0.87	0.67 - 0.93	0.18 - 0.47	0.67 - 0.93	0.77	0.55 - 0.94	0.80 - 0.95	0.35 - 0.56	0.80 - 0.95
	2	0.68	0.22 - 0.95	0.74 - 0.96	0.25 - 0.50	0.74 - 0.96	0.77	0.52 - 0.94	0.82 - 0.97	0.36 - 0.57	0.82 - 0.97



	3	0.62	0.21 - 0.90	0.70 - 0.93	0.20 - 0.50	0.70 - 0.93	0.85	0.58 - 1.02	0.84 - 0.99	0.41 - 0.66	0.84 - 0.99
	4	0.62	0.25 - 0.83	0.68 - 0.94	0.22 - 0.49	0.68 - 0.94	0.72	0.50 - 0.88	0.75 - 0.97	0.33 - 0.56	0.75 - 0.97
	5	0.67	0.30 - 0.89	0.75 - 0.95	0.22 - 0.54	0.75 - 0.95	0.75	0.52 - 0.94	0.80 - 0.98	0.35 - 0.59	0.80 - 0.98
	6	0.65	0.33 - 0.90	0.76 - 0.96	0.26 - 0.53	0.76 - 0.96	0.77	0.50 - 0.95	0.84 - 0.99	0.41 - 0.63	0.84 - 0.99
	7	0.63	0.25 - 0.86	0.74 - 0.95	0.18 - 0.48	0.74 - 0.95	0.80	0.57 - 0.93	0.84 - 0.98	0.43 - 0.63	0.84 - 0.98
	8	0.64	0.26 - 0.88	0.74 - 0.96	0.23 - 0.49	0.74 - 0.96	0.74	0.45 - 0.93	0.82 - 0.98	0.35 - 0.56	0.82 - 0.98
	9	0.70	0.33 - 0.92	0.80 - 0.97	0.29 - 0.58	0.80 - 0.97	0.86	0.65 - 1.01	0.89 - 1.01	0.53 - 0.73	0.89 - 1.01
	10	0.64	0.19 - 0.89	0.72 - 0.94	0.22 - 0.50	0.72 - 0.94	0.75	0.39 - 0.94	0.81 - 0.98	0.33 - 0.58	0.81 - 0.98
	1	0.63	0.34 - 0.80	0.73 - 0.94	0.22 - 0.46	0.73 - 0.94	0.77	0.58 - 0.91	0.84 - 0.99	0.38 - 0.58	0.84 - 0.99
	2	0.68	0.46 - 0.83	0.77 - 0.94	0.26 - 0.54	0.77 - 0.94	0.77	0.58 - 0.89	0.82 - 0.96	0.39 - 0.58	0.82 - 0.96
	3	0.62	0.33 - 0.80	0.70 - 0.92	0.24 - 0.46	0.70 - 0.92	0.85	0.69 - 0.96	0.89 - 1.00	0.49 - 0.68	0.89 - 1.00
	4	0.62	0.35 - 0.80	0.72 - 0.95	0.21 - 0.46	0.72 - 0.95	0.72	0.49 - 0.88	0.81 - 0.99	0.34 - 0.53	0.81 - 0.99
250	5	0.62	0.37 - 0.84	0.70 - 0.90	0.21 - 0.49	0.70 - 0.90	0.75	0.56 - 0.89	0.77 - 0.92	0.30 - 0.50	0.77 - 0.92
	6	0.65	0.40 - 0.84	0.76 - 0.96	0.26 - 0.49	0.76 - 0.96	0.77	0.56 - 0.89	0.85 - 0.99	0.40 - 0.58	0.85 - 0.99
	7	0.63	0.36 - 0.82	0.73 - 0.92	0.21 - 0.48	0.73 - 0.92	0.80	0.60 - 0.91	0.83 - 0.95	0.37 - 0.62	0.83 - 0.95
	8	0.64	0.40 - 0.81	0.71 - 0.93	0.23 - 0.49	0.71 - 0.93	0.74	0.54 - 0.89	0.77 - 0.95	0.35 - 0.54	0.77 - 0.95
	9	0.70	0.47 - 0.85	0.80 - 0.95	0.28 - 0.52	0.80 - 0.95	0.86	0.70 - 0.98	0.90 - 1.00	0.52 - 0.70	0.90 - 1.00
	10	0.64	0.38 - 0.83	0.66 - 0.88	0.16 - 0.43	0.66 - 0.88	0.75	0.50 - 0.89	0.72 - 0.88	0.27 - 0.48	0.72 - 0.88
	1	0.63	0.38 - 0.77	0.73 - 0.94	0.22 - 0.47	0.77 - 0.92	0.77	0.59 - 0.88	0.84 - 0.99	0.38 - 0.58	0.86 - 0.98
	2	0.68	0.46 - 0.84	0.77 - 0.94	0.24 - 0.46	0.73 - 0.89	0.77	0.58 - 0.89	0.82 - 0.96	0.35 - 0.51	0.79 - 0.92
	3	0.62	0.40 - 0.78	0.70 - 0.92	0.21 - 0.44	0.72 - 0.90	0.85	0.69 - 0.96	0.89 - 1.00	0.45 - 0.63	0.88 - 0.98
	4	0.62	0.37 - 0.74	0.72 - 0.95	0.19 - 0.41	0.67 - 0.87	0.72	0.51 - 0.85	0.81 - 0.99	0.27 - 0.47	0.75 - 0.91
	5	0.67	0.41 - 0.84	0.70 - 0.90	0.26 - 0.51	0.77 - 0.93	0.75	0.56 - 0.89	0.77 - 0.92	0.35 - 0.56	0.84 - 0.96
350	6	0.65	0.34 - 0.75	0.76 - 0.96	0.23 - 0.47	0.74 - 0.92	0.77	0.54 - 0.86	0.85 - 0.99	0.36 - 0.54	0.82 - 0.94
	7	0.63	0.40 - 0.79	0.73 - 0.92	0.27 - 0.50	0.76 - 0.92	0.80	0.64 - 0.92	0.83 - 0.95	0.45 - 0.62	0.87 - 0.97
	8	0.64	0.38 - 0.75	0.71 - 0.93	0.26 - 0.47	0.76 - 0.92	0.74	0.57 - 0.84	0.77 - 0.95	0.38 - 0.54	0.84 - 0.95
	9	0.70	0.46 - 0.81	0.80 - 0.95	0.32 - 0.53	0.81 - 0.95	0.86	0.71 - 0.93	0.90 - 1.00	0.54 - 0.70	0.90 - 1.00
	10	0.64	0.40 - 0.78	0.66 - 0.88	0.25 - 0.50	0.74 - 0.93	0.75	0.56 - 0.87	0.72 - 0.88	0.38 - 0.57	0.84 - 0.97



- <sup>b</sup> Number of observations
- <sup>c</sup> True value of loadings
- <sup>d</sup> Model based estimated loadings
- <sup>e</sup> Single selected day
- <sup>f</sup> Weekly item average
- <sup>g</sup> Within-individual analysis
- <sup>h</sup> Between-individual analysis



### 6.1.3 ICC results

ICC coefficients were also computed with the simulated datasets as a verification step regarding the utility or relevance of both sources of variance in explaining variability in the data. The ICC range for each of the items is presented in Table 6.2. Across both scenarios, none was below the cut-offs for inferring an absence of variance (i.e.,  $<0.05$  and  $>0.95$ , respectively) at the within- (i.e.,  $ICC < 0.05$ ) or between-individual levels (i.e.,  $ICC > 0.95$ ), indicating that a multilevel analysis strategy was appropriate. For the scenario 1 the values were lower in comparison with scenario 2 (i.e., 0.10-0.46 across item 1-20 for scenario 1 with  $n=100$  and 0.10-0.52 across item 1-20 for scenario 2,  $n=100$ ), indicating that the between-variability for the latter explained a greater proportion of the variability of the items. Additionally, for both scenarios across all sample sizes, there was not a very distinct difference between the ICC range and the corresponding average value.



Table 6.2: Range and average value of the intraclass correlation coefficient for each item across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

Scenario	Item	n <sup>a</sup> =100		n <sup>a</sup> =150		n <sup>a</sup> =200		n <sup>a</sup> =250		n <sup>a</sup> =350	
		ICC <sup>b</sup> Range	ICC <sup>b</sup> Average	ICC <sup>b</sup> Range	ICC <sup>b</sup> Average	ICC <sup>b</sup> Range	ICC <sup>b</sup> Average	ICC <sup>b</sup> Range	ICC <sup>b</sup> Average	ICC <sup>b</sup> Range	ICC <sup>b</sup> Average
1	Item 1	0.20 - 0.46	0.33	0.30 - 0.44	0.37	0.30 - 0.44	0.36	0.30 - 0.42	0.36	0.30 - 0.42	0.37
	Item 2	0.13 - 0.32	0.23	0.19 - 0.36	0.27	0.23 - 0.37	0.29	0.23 - 0.36	0.29	0.20 - 0.32	0.25
	Item 3	0.14 - 0.33	0.23	0.10 - 0.29	0.19	0.18 - 0.33	0.25	0.19 - 0.30	0.25	0.18 - 0.29	0.23
	Item 4	0.13 - 0.31	0.22	0.17 - 0.33	0.24	0.17 - 0.33	0.24	0.16 - 0.30	0.23	0.15 - 0.30	0.22
	Item 5	0.17 - 0.35	0.25	0.20 - 0.36	0.27	0.22 - 0.35	0.27	0.19 - 0.32	0.25	0.23 - 0.34	0.28
	Item 6	0.14 - 0.34	0.24	0.17 - 0.34	0.24	0.21 - 0.36	0.28	0.22 - 0.35	0.28	0.20 - 0.31	0.25
	Item 7	0.12 - 0.32	0.22	0.15 - 0.32	0.23	0.18 - 0.33	0.25	0.18 - 0.31	0.24	0.21 - 0.32	0.26
	Item 8	0.15 - 0.33	0.23	0.13 - 0.32	0.22	0.20 - 0.34	0.26	0.20 - 0.35	0.26	0.21 - 0.32	0.26
	Item 9	0.19 - 0.40	0.29	0.23 - 0.37	0.30	0.25 - 0.38	0.31	0.26 - 0.40	0.31	0.27 - 0.39	0.32
	Item 10	0.11 - 0.32	0.21	0.14 - 0.31	0.22	0.19 - 0.34	0.26	0.18 - 0.32	0.24	0.21 - 0.32	0.26
	Item 11	0.13 - 0.28	0.11	0.17 - 0.31	0.23	0.15 - 0.28	0.21	0.14 - 0.29	0.21	0.14 - 0.25	0.19
	Item 12	0.14 - 0.33	0.22	0.22 - 0.35	0.28	0.20 - 0.34	0.26	0.21 - 0.33	0.26	0.24 - 0.35	0.29
	Item 13	0.10 - 0.30	0.20	0.15 - 0.34	0.24	0.20 - 0.33	0.25	0.19 - 0.35	0.27	0.19 - 0.31	0.25
	Item 14	0.14 - 0.32	0.22	0.16 - 0.33	0.24	0.19 - 0.32	0.25	0.18 - 0.31	0.24	0.21 - 0.31	0.26
	Item 15	0.09 - 0.27	0.10	0.21 - 0.35	0.28	0.20 - 0.35	0.28	0.25 - 0.38	0.31	0.20 - 0.32	0.26
	Item 16	0.16 - 0.38	0.25	0.18 - 0.33	0.25	0.19 - 0.34	0.25	0.18 - 0.33	0.25	0.20 - 0.30	0.25
	Item 17	0.14 - 0.35	0.23	0.14 - 0.31	0.21	0.20 - 0.32	0.25	0.19 - 0.32	0.26	0.21 - 0.31	0.26
	Item 18	0.21 - 0.42	0.31	0.22 - 0.39	0.30	0.24 - 0.37	0.30	0.26 - 0.38	0.31	0.25 - 0.36	0.30
	Item 19	0.25 - 0.45	0.33	0.26 - 0.41	0.33	0.26 - 0.40	0.32	0.25 - 0.38	0.31	0.26 - 0.37	0.32
	Item 20	0.17 - 0.37	0.26	0.23 - 0.40	0.31	0.21 - 0.34	0.27	0.22 - 0.34	0.28	0.22 - 0.33	0.27
2	Item 1	0.20 - 0.46	0.33	0.30 - 0.44	0.37	0.30 - 0.44	0.36	0.30 - 0.42	0.36	0.30 - 0.42	0.37
	Item 2	0.20 - 0.41	0.30	0.30 - 0.43	0.36	0.30 - 0.45	0.38	0.30 - 0.44	0.37	0.30 - 0.40	0.33
	Item 3	0.30 - 0.52	0.42	0.30 - 0.48	0.38	0.40 - 0.50	0.43	0.40 - 0.51	0.45	0.40 - 0.47	0.41
	Item 4	0.20 - 0.40	0.30	0.20 - 0.41	0.33	0.30 - 0.39	0.32	0.30 - 0.38	0.32	0.20 - 0.38	0.32
	Item 5	0.20 - 0.42	0.32	0.30 - 0.43	0.35	0.30 - 0.41	0.34	0.30 - 0.39	0.32	0.30 - 0.41	0.36



---

Item 6	0.30 - 0.43	0.34	0.30 - 0.43	0.35	0.30 - 0.45	0.38	0.30 - 0.44	0.38	0.30 - 0.41	0.36
Item 7	0.30 - 0.45	0.35	0.30 - 0.46	0.38	0.40 - 0.49	0.41	0.30 - 0.44	0.38	0.40 - 0.45	0.40
Item 8	0.20 - 0.41	0.32	0.20 - 0.38	0.30	0.30 - 0.41	0.34	0.30 - 0.42	0.34	0.30 - 0.39	0.33
Item 9	0.40 - 0.52	0.44	0.40 - 0.51	0.44	0.40 - 0.54	0.49	0.40 - 0.53	0.47	0.40 - 0.52	0.47
Item 10	0.20 - 0.40	0.29	0.20 - 0.39	0.30	0.30 - 0.44	0.36	0.30 - 0.42	0.32	0.30 - 0.41	0.35
Item 11	0.10 - 0.32	0.20	0.30 - 0.45	0.36	0.30 - 0.42	0.35	0.30 - 0.43	0.35	0.20 - 0.36	0.30
Item 12	0.20 - 0.41	0.31	0.30 - 0.46	0.39	0.30 - 0.44	0.37	0.30 - 0.41	0.35	0.40 - 0.47	0.41
Item 13	0.10 - 0.36	0.24	0.20 - 0.39	0.30	0.30 - 0.38	0.31	0.30 - 0.39	0.33	0.20 - 0.36	0.30
Item 14	0.20 - 0.33	0.24	0.20 - 0.36	0.27	0.20 - 0.35	0.28	0.20 - 0.33	0.27	0.20 - 0.34	0.29
Item 15	0.20 - 0.37	0.28	0.30 - 0.48	0.40	0.30 - 0.49	0.42	0.40 - 0.53	0.48	0.30 - 0.45	0.39
Item 16	0.20 - 0.43	0.32	0.20 - 0.40	0.31	0.30 - 0.40	0.33	0.30 - 0.40	0.33	0.30 - 0.37	0.32
Item 17	0.20 - 0.42	0.32	0.20 - 0.38	0.29	0.30 - 0.40	0.34	0.30 - 0.41	0.35	0.30 - 0.40	0.35
Item 18	0.30 - 0.46	0.36	0.30 - 0.42	0.34	0.30 - 0.41	0.35	0.30 - 0.43	0.37	0.30 - 0.39	0.35
Item 19	0.30 - 0.50	0.39	0.30 - 0.47	0.38	0.30 - 0.44	0.37	0.30 - 0.43	0.36	0.30 - 0.42	0.37
Item 20	0.30 - 0.49	0.38	0.40 - 0.50	0.43	0.30 - 0.45	0.37	0.30 - 0.46	0.40	0.30 - 0.45	0.39

---

<sup>a</sup> Number of individuals

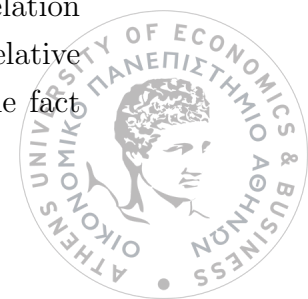
<sup>b</sup> Intraclass correlation coefficient



### 6.1.4 Estimated, true and observed inter-item correlation

Table 6.3 presents the observed inter-item correlations alongside with the correlation values under the hypothesized model (true correlation in this simulated data framework) and the EFA-estimated correlation values. In an examination of the different simulated and estimated correlation matrices, it was seen that the weekly item average approach produced higher observed correlations (and of a greater range) in comparison with the single selected day approach. This is consistent with the findings established in earlier stages of the analysis regarding the greater consistency in factor identification observed with the weekly item average approach across both scenarios and various sample sizes. As expected, scenario 2 produced higher inter item correlations for both single selected day and weekly item average approach in comparison with scenario 1. For instance, single selected day approach in scenario 1 produced observed correlation that ranged from -0.08 to 0.77 with  $n=100$  and for scenario 2 this correlation ranged from 0.10 to 0.84). In accordance with expectations, these patterns were also observed in EFA-inter item correlation estimates which produced slightly lower values compared to the observed ones (i.e.,  $r=0.56-0.86$  for the estimated correlation of weekly item average approach and  $r=0.52-0.88$  for the observed correlation of weekly item average approach, scenario 1,  $n=100$ ). Additionally, as the sample size increased the range for both model based estimates and observed correlations, became narrower. For the single-selected day, the range became more close to the range of the true inter item correlation. This could explain also the increased percentage of correct identification of the number of factor for the single selected day when the number of individual was greater than 100. The results demonstrated that the weekly item average approach systematically produced higher correlations than the true values of the correlation matrix under both scenarios and all selected sample sizes. Additionally in the single selected day approach, the range of the observed inter-item correlations was quite similar to the range of the true correlation matrix as was previously flagged. Similar findings are observed when comparing the EFA-estimated correlation ranges with the true correlation.

To aid the interpretation of the results, Figure 6.3 was produced, which visualizes the range (defined as the difference of the maximum and minimum value) of the estimated and observed correlation within each factor for both single selected and weekly item average approach across both scenarios. Note that the current figure correspond to the first factor and that other the figures for the other factors are provided in Appendix B. Based on this figure it is evident that there is a difference between estimated and observed correlation as was previously flagged as the range of the estimated correlation is slightly lower. Additionally, the increased correlation of weekly item average relative to the single selected day across both scenarios is well depicted, along with the fact



that the increase of sample size resulted in a less narrow range of the correlation across both approaches.



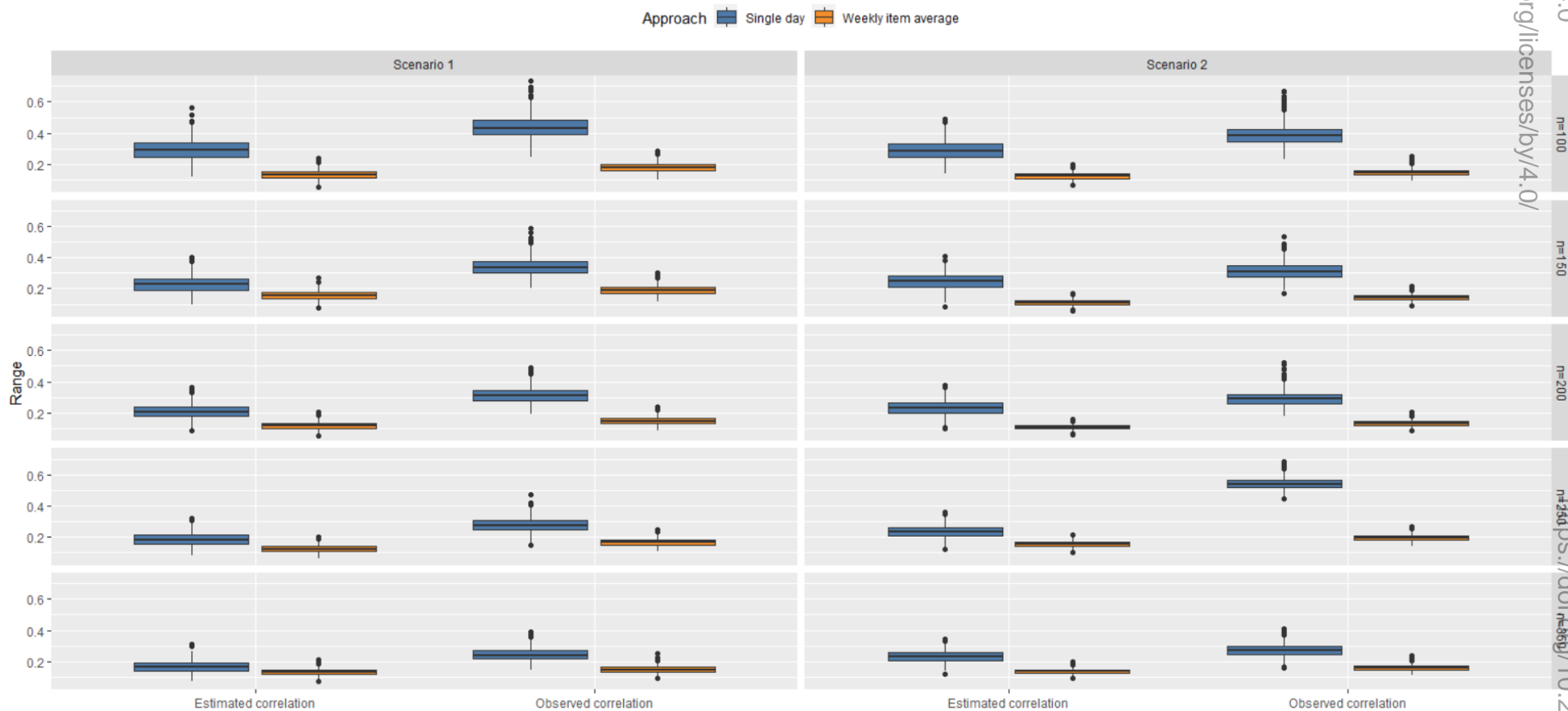
Table 6.3: True correlation parameters, observed inter-item correlations, and EFA-estimated correlations for the single selected day and weekly item average approaches with the 1,000 simulated datasets 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

n <sup>b</sup>	Factor	Scenario 1					Scenario 2				
		$R_T^c$	$R_M^d$	$R_O^e$	$R_T^c$	$R_M^d$	$R_O^e$	$R_T^c$	$R_M^d$	$R_O^e$	
		SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>
100	1	0.39 - 0.48	0.04 - 0.71	0.56 - 0.86	-0.08 - 0.78	0.52 - 0.88	0.55 - 0.74	0.25 - 0.83	0.69 - 0.93	0.10 - 0.84	0.67 - 0.94
	2	0.41 - 0.49	-0.06 - 0.75	0.38 - 0.83	-0.06 - 0.75	0.30 - 0.83	0.51 - 0.71	0.21 - 0.87	0.55 - 0.89	-0.09 - 1.00	0.45 - 0.89
	3	0.39 - 0.49	0.07 - 0.82	0.59 - 0.87	-0.02 - 0.81	0.51 - 0.88	0.53 - 0.60	0.30 - 0.87	0.68 - 0.92	0.23 - 0.88	0.63 - 0.92
150	1	0.39 - 0.48	0.19 - 0.69	0.54 - 0.84	0.07 - 0.74	0.49 - 0.86	0.55 - 0.74	0.38 - 0.86	0.71 - 0.91	0.27 - 0.87	0.66 - 0.92
	2	0.41 - 0.49	0.18 - 0.75	0.59 - 0.84	0.11 - 0.74	0.54 - 0.84	0.51 - 0.71	0.30 - 0.87	0.64 - 0.91	0.15 - 0.87	0.60 - 0.92
	3	0.39 - 0.49	0.22 - 0.79	0.59 - 0.87	0.13 - 0.79	0.57 - 0.88	0.53 - 0.60	0.35 - 0.83	0.69 - 0.90	0.29 - 0.84	0.69 - 0.90
200	1	0.39 - 0.48	0.19 - 0.66	0.59 - 0.83	0.09 - 0.71	0.57 - 0.84	0.55 - 0.74	0.38 - 0.82	0.73 - 0.92	0.29 - 0.84	0.70 - 0.93
	2	0.41 - 0.49	0.19 - 0.63	0.60 - 0.83	0.10 - 0.65	0.57 - 0.83	0.51 - 0.71	0.33 - 0.80	0.70 - 0.90	0.21 - 0.82	0.69 - 0.91
	3	0.39 - 0.49	0.22 - 0.68	0.65 - 0.84	0.18 - 0.70	0.65 - 0.85	0.53 - 0.60	0.39 - 0.77	0.75 - 0.88	0.37 - 0.79	0.74 - 0.88
250	1	0.39 - 0.48	0.23 - 0.62	0.58 - 0.81	0.11 - 0.68	0.53 - 0.83	0.55 - 0.74	0.38 - 0.80	0.68 - 0.91	0.31 - 1.00	0.63 - 0.93
	2	0.41 - 0.49	0.22 - 0.63	0.54 - 0.82	0.13 - 0.66	0.51 - 0.82	0.51 - 0.71	0.33 - 0.79	0.65 - 0.89	0.24 - 0.79	0.60 - 0.90
	3	0.39 - 0.49	0.27 - 0.64	0.64 - 0.84	0.20 - 0.67	0.61 - 0.84	0.53 - 0.60	0.38 - 0.75	0.75 - 0.88	0.36 - 0.75	0.74 - 0.88
350	1	0.39 - 0.48	0.21 - 0.58	0.58 - 0.82	0.15 - 0.60	0.56 - 0.83	0.55 - 0.74	0.39 - 0.80	0.69 - 0.91	0.34 - 0.83	0.66 - 0.91
	2	0.41 - 0.49	0.23 - 0.61	0.56 - 0.80	0.20 - 0.64	0.55 - 0.80	0.51 - 0.71	0.37 - 0.80	0.67 - 0.89	0.31 - 0.81	0.65 - 0.89
	3	0.39 - 0.49	0.25 - 0.61	0.65 - 0.82	0.20 - 0.63	0.64 - 0.83	0.53 - 0.60	0.41 - 0.71	0.75 - 0.87	0.37 - 0.72	0.73 - 0.87

<sup>a</sup> Exploratory Factor Analysis  
<sup>b</sup> Number of individuals  
<sup>c</sup> Range of the true correlation between items  
<sup>d</sup> Range of the observed correlation between items  
<sup>e</sup> Range of the model-based estimated correlation between items under the EFA model  
<sup>f</sup> Single selected day  
<sup>g</sup> Weekly item average



Figure 6.3: Boxplot of the range of inter item observed and EFA-estimated correlation within the second factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis Single day=Single selected day approach; Item average= Weekly item average; a: Slope parameter from a multidimensional graded response model.



### 6.1.5 Overall factor loading bias and mean squared error

Overall bias and MSE were calculated for the weekly item average and single selected day approaches to quantify the distance of the estimated loadings under EFA and the true parameters (see Figure 6.4 for bias and 6.5 for MSE). The weekly item average approach systematically produced higher overall bias, whereas the single selected day approach produced closer values to zero. Scenario 2 where the true loadings were higher than scenario 1 resulted in less bias. This difference was more apparent for the weekly item average approach, where its values were more closely aligned with the corresponding bias of the single selected day approach. Similarly, the overall MSE of the loadings was lower for the single selected day than the corresponding MSE of the weekly item average approach, with similar patterns being observed as before in regard to the differences between scenario 2 and scenario 1. The prior analysis of the factor loading ranges demonstrated that the weekly item average approach systematically provided higher range than the corresponding range of the factor loading parameters, indicating that this approach is prone to overestimation compared to single selected day which did not show any systematic pattern due to its wide range.



Figure 6.4: Overall absolute average bias of the estimated loadings of items within the first factor for the EFA model with 1,000 simulated datasets for 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Item average=Weekly item average; a: Slope parameter from a multidimensional graded response model.

120

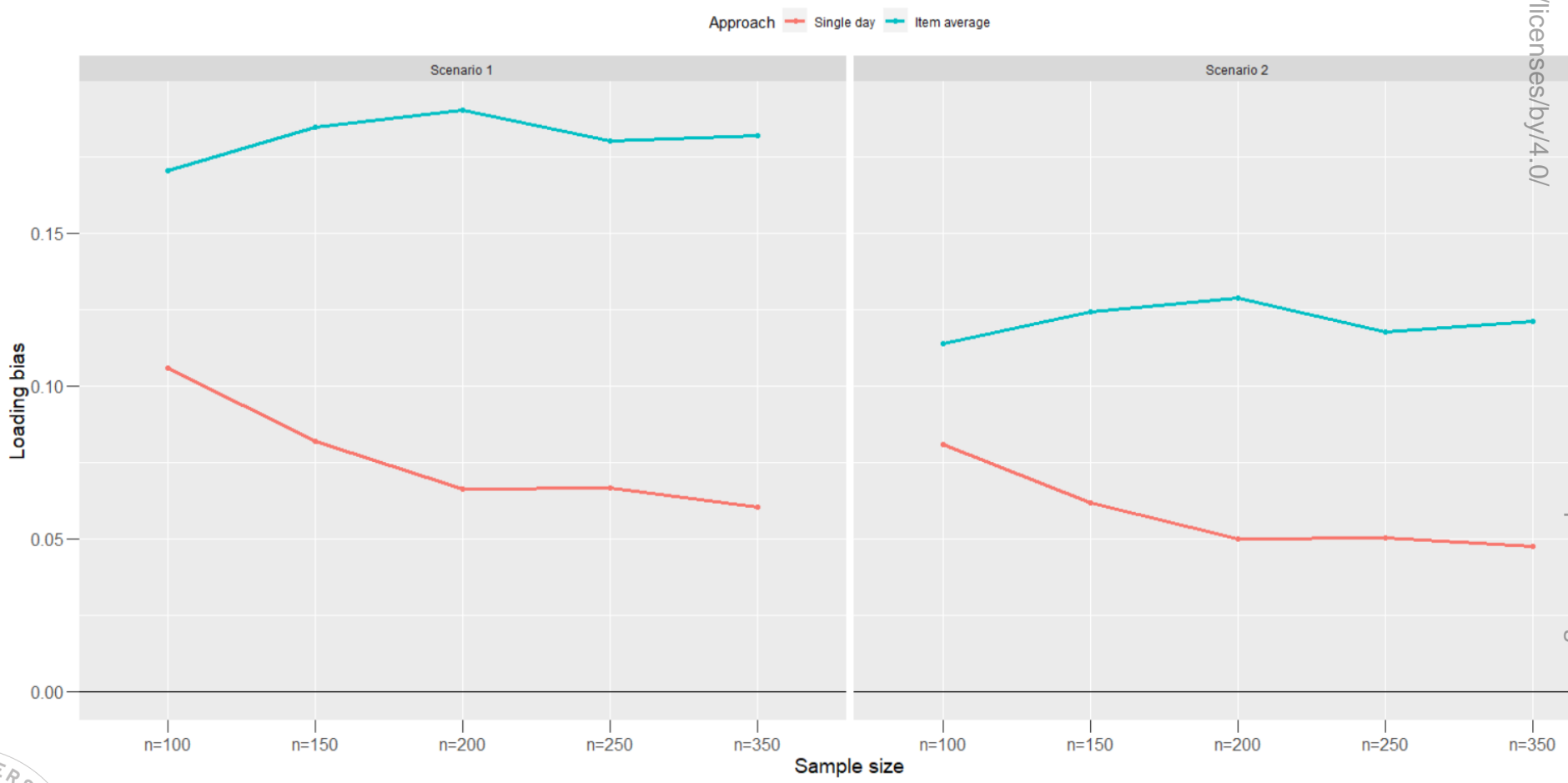
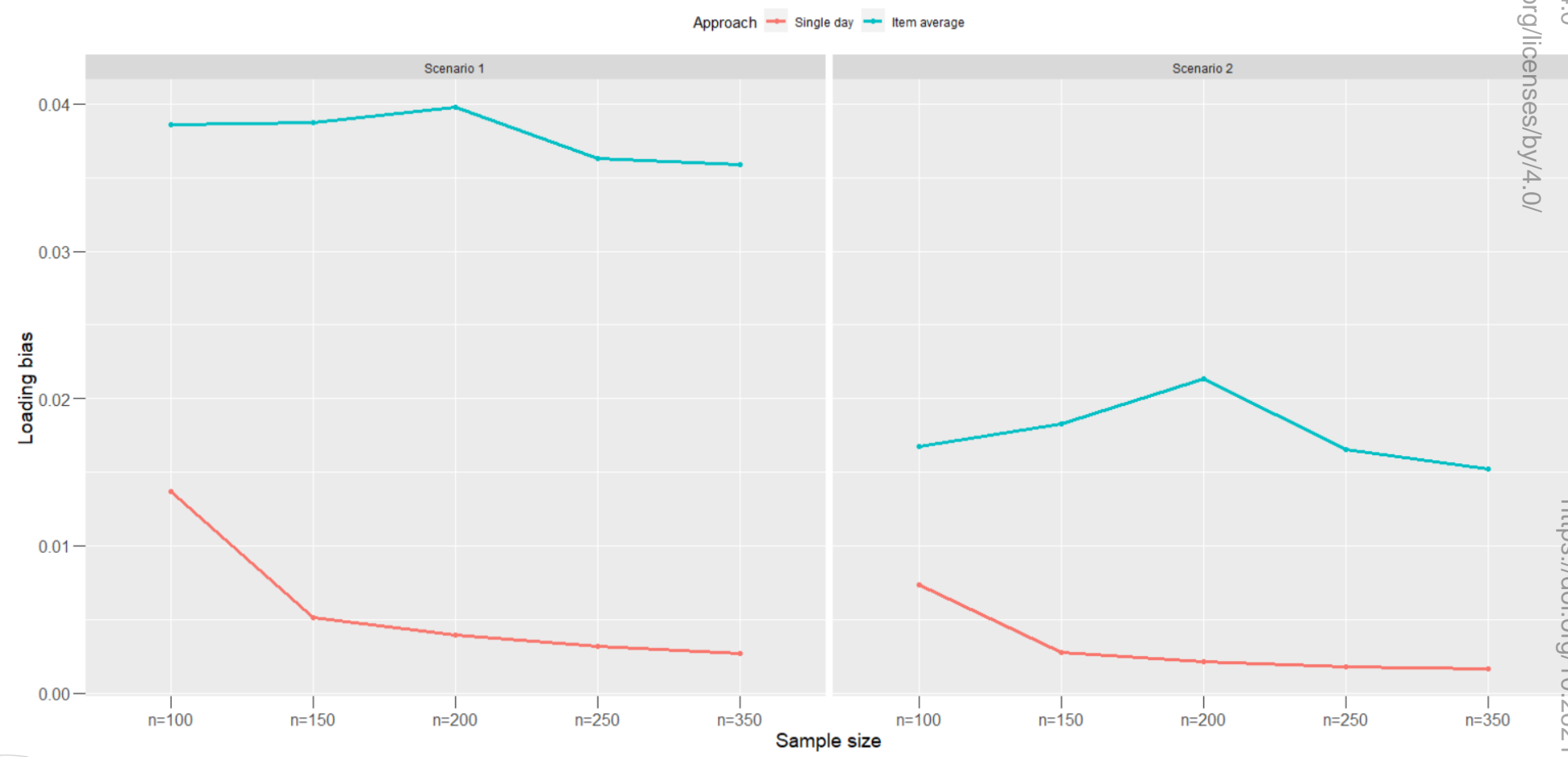


Figure 6.5: Overall MSE of the estimated loadings of items within the first factor for the CFA model with 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Item average=Weekly item average; MSE: Mean square error; a: Slope parameter from a multidimensional graded response model.



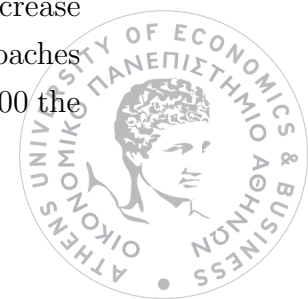
### 6.1.6 Goodness of fit measures

Summary statistics for the goodness of fit results across all iterations of the simulated data across the two scenarios are presented for each of the EFA modelling approaches in Figure 6.6, 6.7 and 6.8.

None of the modelling strategies evidenced poor fit when evaluated via the SRMR. Across all goodness of fit measures, all modelling approaches had almost invariably good or acceptable fit. TLI seemed to be the only one suggesting a bad model fit in some occasions, especially for within-individual analysis model, with the corresponding percentage being decreased as the sample size increased for scenario 1 whereas in scenario 2 this percentage remained at the same levels. Generally this discrepancy between TLI and the other goodness of fit measures is evident in the literature (Bentler, 1990; Kline, 2015) as TLI can suggest a bad fit when evaluating a model even when the other measures suggest a good fit. What is more, when the sample size became greater than 100 the corresponding percentage of bad fit was reduced to zero for the single day, weekly item average approach and between-individual analysis model. An important thing to note was the higher percentage of bad fit for scenario 1 in comparison with scenario 2 when the number of individuals was equal to 100. The reason for this occurrence might be that the true indicator factor patterns were quite weaker for scenario 1 and the observed indicator patterns of the within-individual models were even weaker as showed in the previous results, which in some iterations might have led to a really weak factor structure in contrast with the weekly item average, single selected day and between-individual analysis where the association strength among factors and items was higher. CFI also seemed to suggest a bad fit for the single day approach and within-individual analysis, but the percentage was lower than the case of TLI. Generally, the percentage of bad fit across all approaches was quite low.

When looking the percentage of acceptable fit, the RMSEA demonstrated the highest percentage of acceptable fit for the weekly item average and between-individual model, SRMR for the single selected day and CFI and TLI for the within-individual analysis. Interestingly enough, for the single selected day approach the corresponding percentage for SRMR was quite high for both scenarios, which was then significantly decreased with the increase of the sample size. As expected, RMSEA was sample size dependent as the percentage of acceptable fit when the number of individual was 100 for both scenarios was high (i.e, see between-individual model) and then this percentage was significantly decreased resulting in higher percentages of good fit.

A final inspection was dedicated also to the percentage of good fit, where the increase of the sample size led to an increase of the percentage of good fit across all approaches and scenarios. A notable difference between scenario 1 and 2 for was that for  $n=100$  the



percentage of good fit based on RMSEA was higher on scenario 1 compared to scenario 2 for the weekly item average approach and between-individual model. Overall, when  $n=100$  for the single day approach, RMSEA and CFI suggested more often good fit while for the weekly item average approach and between-individual analysis SRMR and CFI showed higher percentages. As for within-individual model, RMSEA and SRMR suggested more frequently good fit. Given the divergence of the estimated correlation for between-individual model and weekly item average approach with the true correlation of the simulated model, TLI and RMSEA seemed to perform better as they seemed to be more conservative.

Consequently, across all the goodness of fit measures TLI was the one who showed higher percentages of poor fit in comparison with the RMSEA, CFI and SRMR. SRMR seemed to show a really high percentage of acceptable fit for the single day ( $n=100$ ) whereas for the other sample sizes good fit was more frequently suggested. RMSEA and CFI almost invariably suggested either acceptable or a good fit across all approaches.



Figure 6.6: Bar plots for the percentage of poor fit (see Chapter 4.8) in EFA of the 1,000 simulated datasets for each of the data handling approaches for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model.

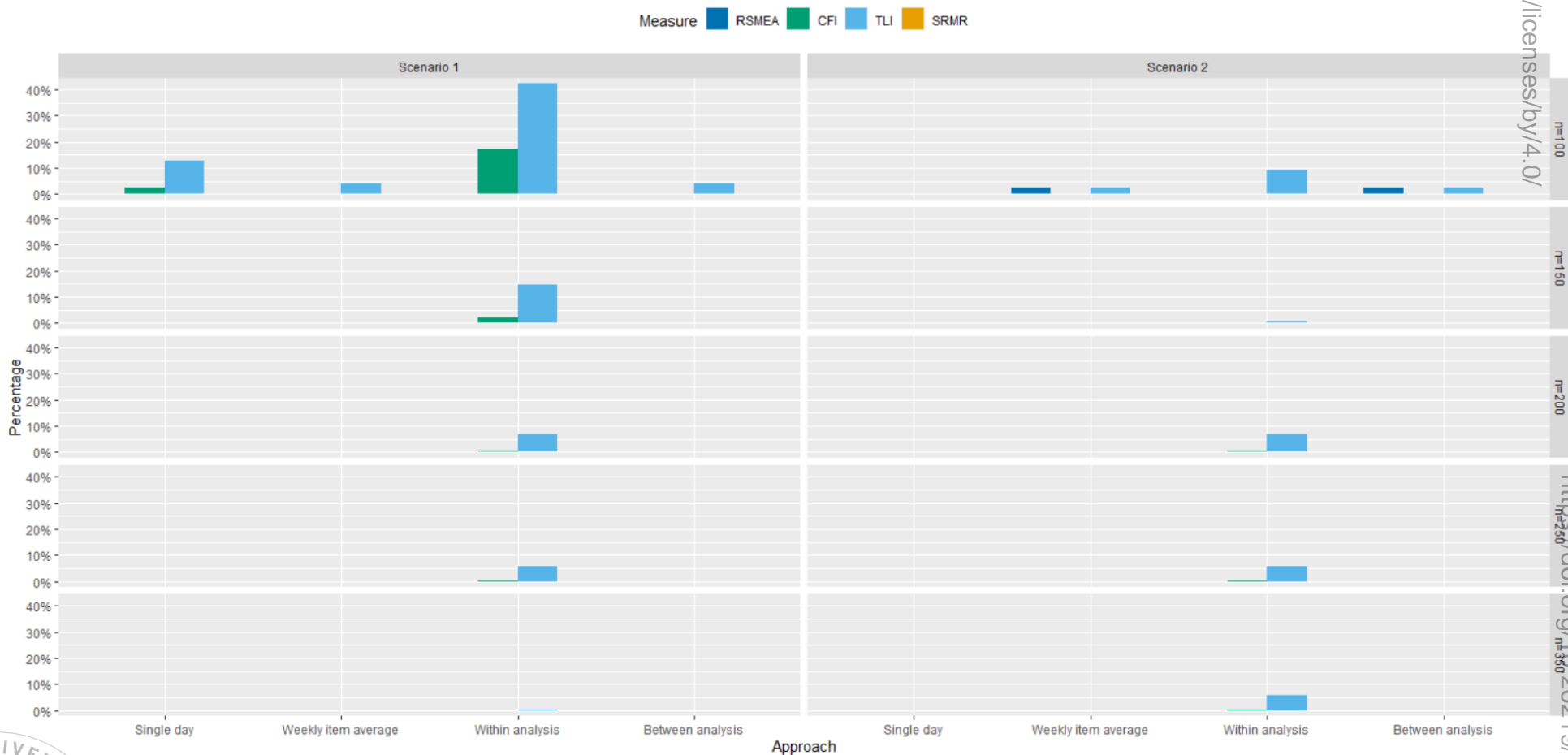


Figure 6.7: Bar plots for the percentage of acceptable fit (see Chapter 4.8) in EFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) based on selected goodness of fit measures. EFA=Exploratory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model.

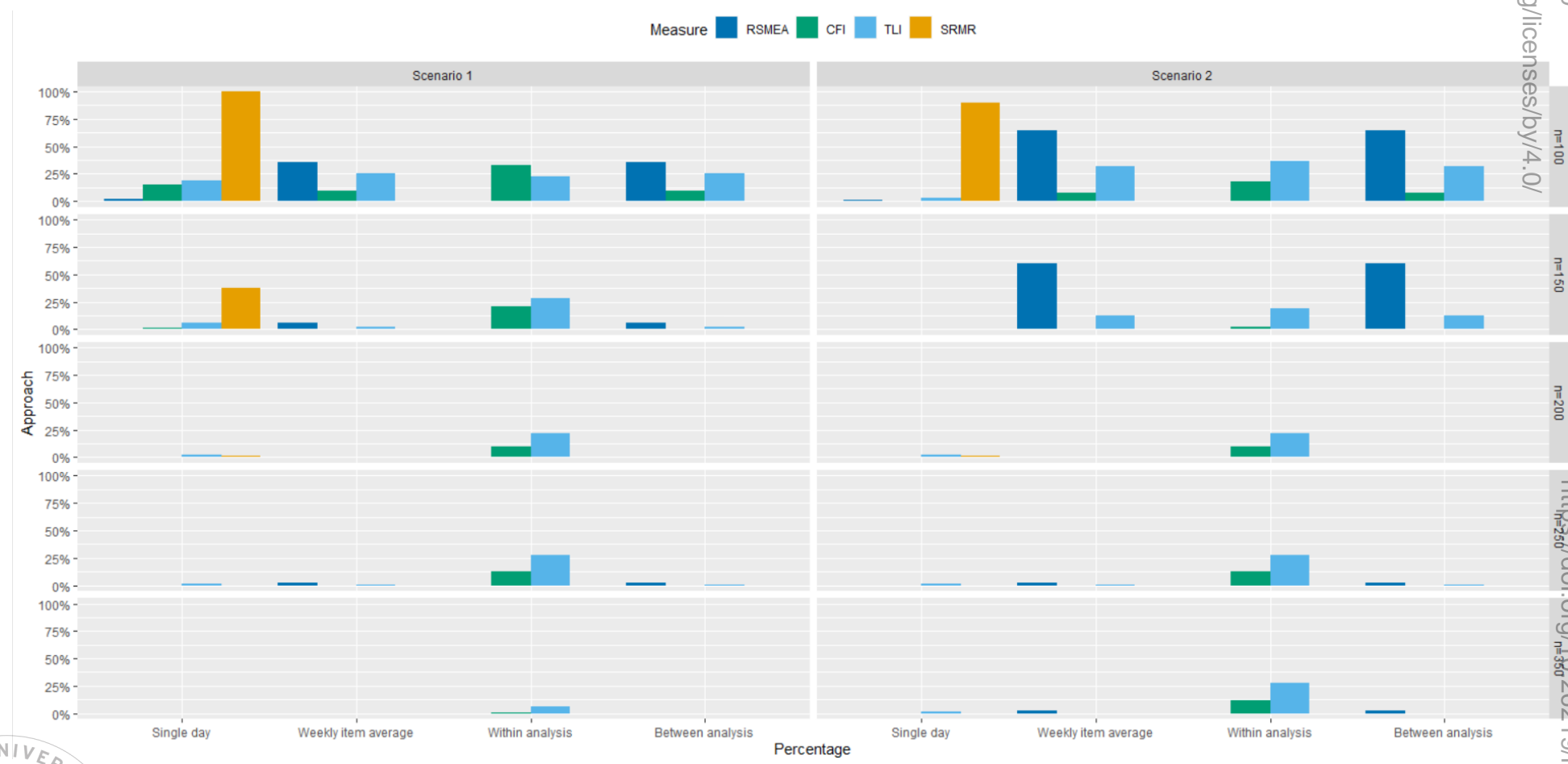
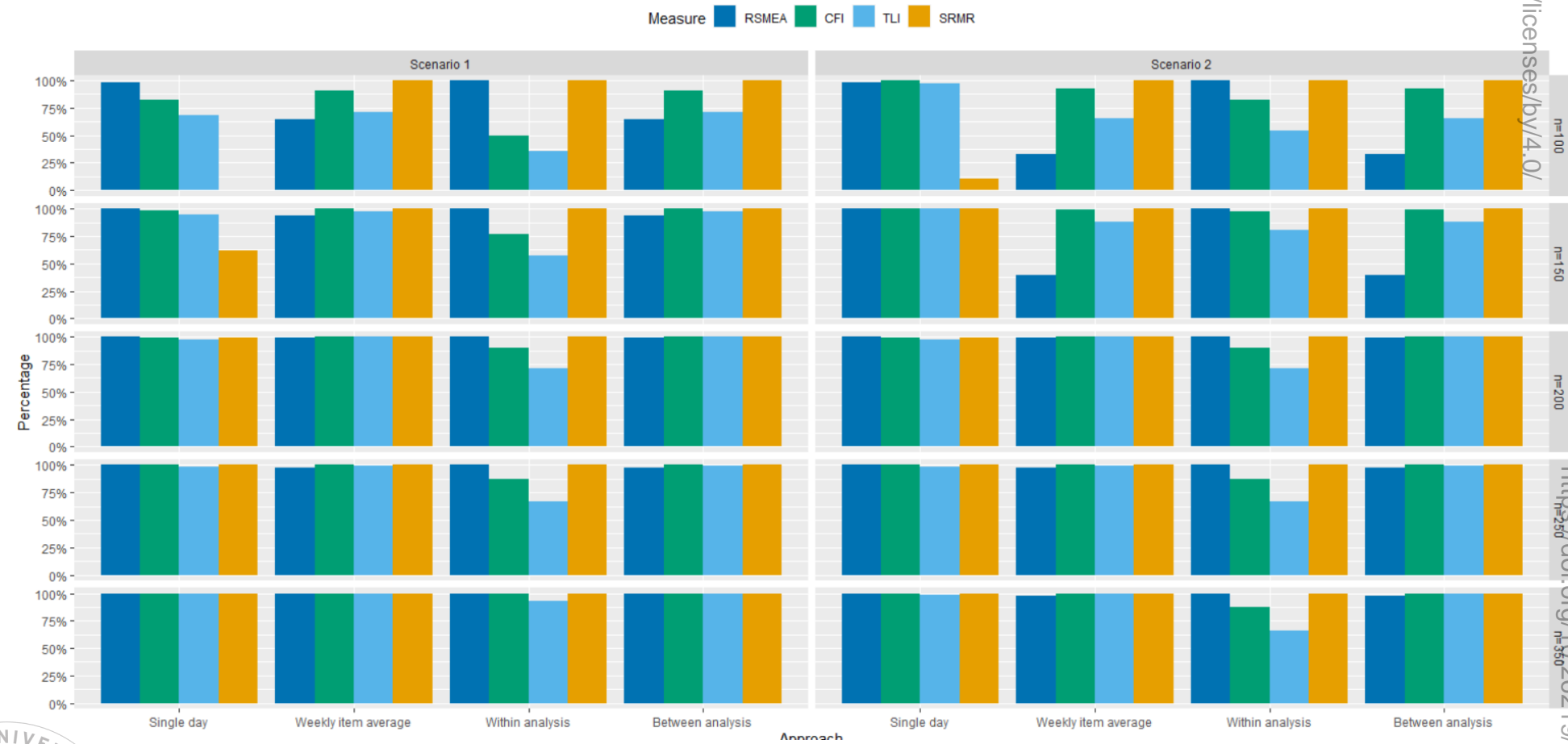


Figure 6.8: Bar plots for the percentage of good fit (see Chapter 4.8) in EFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) based on selected goodness of fit measures. RMSEA=Root mean square error of approximation; EFA=Exploratory factor analysis; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model.



### 6.1.7 Convergence and Heywood cases

A final inspection focussed on possible convergence issues and Heywood cases across all approaches. Based on Table 6.4 there were no such issues apart from the single selected day approach with  $n=100$  where the percentage of convergence was 99.9% and Heywood cases showed frequency of 0.6% for both scenarios. Additionally, within-individual analysis showed 1 occurrence of a Heywood case and no convergence.



Table 6.4: Percentage of convergence and Heywood case for the single selected day, weekly item average, within- and between-individual analysis approaches for the EFA<sup>e</sup>model with 100, 150, 200, 250 and 350 individuals across 1 week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: slope parameter from a multidimensional graded response model

	n	Scenario 1				Scenario 2			
		SD <sup>a</sup>	WIA <sup>b</sup>	WA <sup>c</sup>	BA <sup>d</sup>	SD <sup>a</sup>	WIA <sup>b</sup>	WA <sup>c</sup>	BA <sup>d</sup>
Convergence	100	99.9%	100.0%	100.0%	100.0%	99.0%	100.0%	100.0%	100.0%
	150	100.0%	100.0%	99.9%	100.0%	100.0%	100.0%	99.9%	100.0%
	200	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	250	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	350	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Heywood case	100	0.0%	0.00%	0.00%	0.00%	0.6%	0.00%	0.00%	0.00%
	150	0.00%	0.00%	0.10%	0.00%	0.00%	0.00%	0.00%	0.00%
	200	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	250	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	350	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

- <sup>a</sup> Single selected day
- <sup>b</sup> Weekly item average
- <sup>c</sup> Within-individual analysis
- <sup>d</sup> Between-individual analysis
- <sup>e</sup> Exploratory factor analysis



## 6.2 CFA results

To document an equivalent evaluation under the CFA framework, the following results are presented:

- Factor loading strength and range
- Estimated, true and observed inter-item correlation
- Overall factor loading bias and MSE
- Goodness of fit measures



### 6.2.1 Factor loading strength and range

The range of the loadings for the CFA iterations across all 1,000 simulated datasets were visualized via boxplots (an example of which is provided for the first factor in Figure 6.9, with the corresponding results for the other factors included in Appendix B). The loadings on the within-individual part of the multilevel CFA were weaker than the loadings on the between-individual part of the model. This parallels a finding earlier established in the EFA, although the CFA approach was simultaneously modelling both the within- and between-individual variance, unlike the stratified EFA approach. An interesting result, was the difference between the estimated loadings of the weekly item average approach with the corresponding estimates of the between-individual model, as a result of the different estimation process. Such a difference signified the importance of retrieving estimates for the between-individual model by estimating it simultaneously with the within-individual model. Additional to this, the estimates from the between-individual analysis occasionally included standardized loadings greater than 1 under both scenario 1 and 2, resulting in Heywood cases.

The findings for the other data handling approaches were also similar to what was observed with the earlier series of EFAs.



Figure 6.9: Boxplot of the loadings within the first factor across 1,000 simulated datasets for CFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Between=between-individual analysis; Within=within-individual analysis; a: Slope parameter from a multidimensional graded response model.

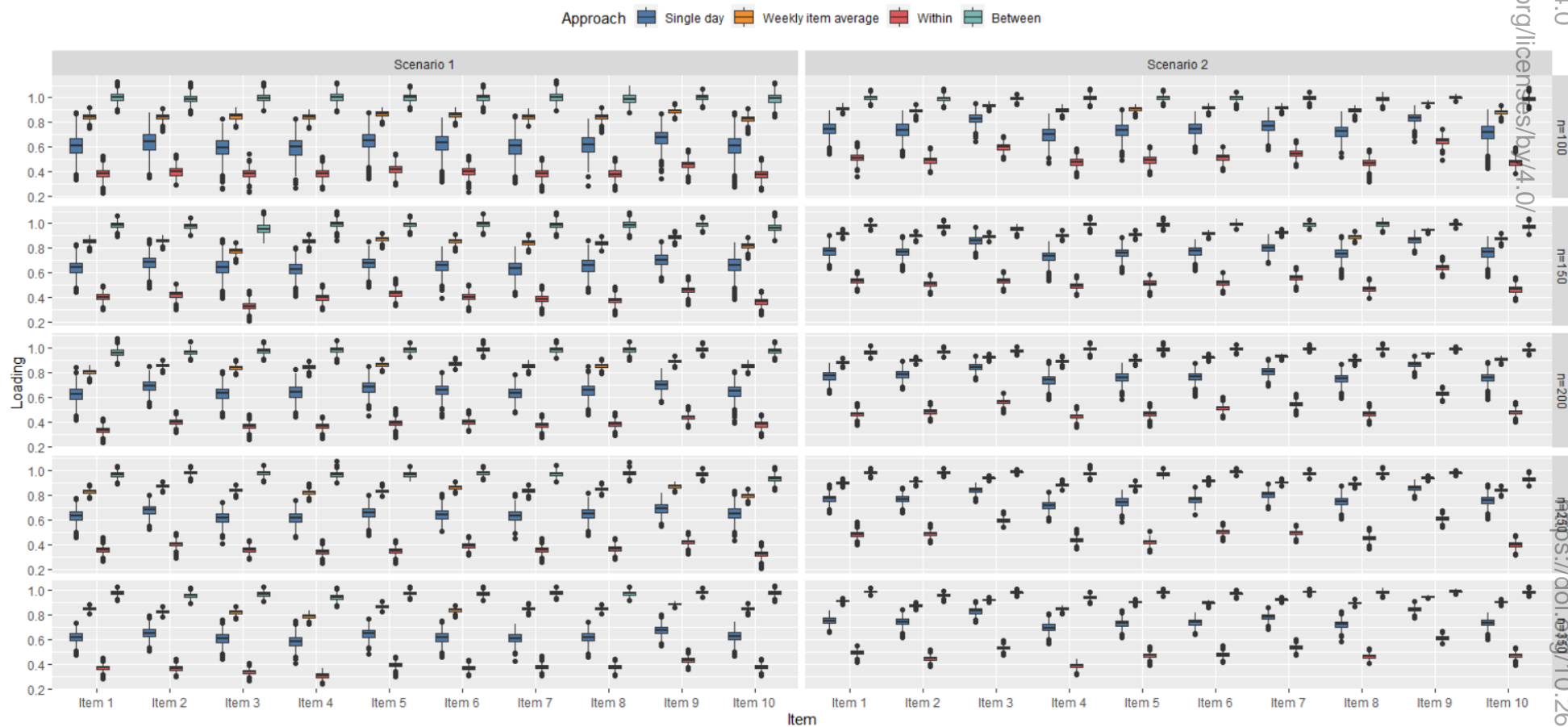


Table 6.5: True loadings parameters, and CFA<sup>a</sup>estimated loadings within factor 1 for the single selected day, weekly item average approaches, multilevel CFA with the 1,000 with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2. a: Slope parameter from a multidimensional graded response model.

n <sup>b</sup>	Item	Scenario 1					Scenario 2				
		$L_T^c$	$L_M^d$				$L_T^c$	$L_M^d$			
			SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>		SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>
100	1	0.63	0.33 - 0.88	0.75 - 0.92	0.22 - 0.53	0.88 - 1.13	0.77	0.54 - 0.90	0.87 - 0.95	0.36 - 0.64	0.94 - 1.06
	2	0.68	0.35 - 0.88	0.72 - 0.91	0.29 - 0.53	0.87 - 1.12	0.77	0.52 - 0.90	0.82 - 0.94	0.39 - 0.59	0.92 - 1.07
	3	0.62	0.26 - 0.83	0.76 - 0.92	0.23 - 0.54	0.89 - 1.12	0.85	0.64 - 0.9	0.89 - 0.97	0.50 - 0.68	0.95 - 1.03
	4	0.62	0.27 - 0.83	0.75 - 0.91	0.26 - 0.52	0.89 - 1.12	0.72	0.47 - 0.87	0.84 - 0.94	0.36 - 0.58	0.93 - 1.07
	5	0.67	0.34 - 0.88	0.78 - 0.92	0.29 - 0.54	0.90 - 1.10	0.75	0.49 - 0.91	0.85 - 0.95	0.38 - 0.60	0.93 - 1.06
	6	0.65	0.32 - 0.84	0.78 - 0.92	0.24 - 0.52	0.89 - 1.11	0.77	0.56 - 0.89	0.86 - 0.95	0.41 - 0.60	0.92 - 1.05
	7	0.63	0.31 - 0.85	0.77 - 0.92	0.24 - 0.51	0.90 - 1.13	0.80	0.57 - 0.92	0.87 - 0.96	0.44 - 0.65	0.93 - 1.05
	8	0.64	0.28 - 0.83	0.72 - 0.92	0.25 - 0.51	0.88 - 1.09	0.74	0.51 - 0.89	0.82 - 0.94	0.31 - 0.57	0.91 - 1.05
	9	0.70	0.34 - 0.87	0.82 - 0.95	0.31 - 0.57	0.92 - 1.07	0.86	0.34 - 0.87	0.82 - 0.95	0.31 - 0.57	0.92 - 1.07
	10	0.64	0.28 - 0.88	0.71 - 0.91	0.25 - 0.51	0.85 - 1.12	0.75	0.43 - 0.91	0.81 - 0.94	0.38 - 0.60	0.91 - 1.08
150	1	0.63	0.44 - 0.83	0.78 - 0.90	0.30 - 0.49	0.89 - 1.07	0.77	0.64 - 0.90	0.87 - 0.95	0.45 - 0.61	0.94 - 1.03
	2	0.68	0.47 - 0.87	0.79 - 0.91	0.30 - 0.51	0.90 - 1.05	0.77	0.61 - 0.90	0.85 - 0.94	0.42 - 0.59	0.92 - 1.03
	3	0.62	0.40 - 0.87	0.68 - 0.85	0.21 - 0.45	0.84 - 1.09	0.85	0.73 - 0.97	0.85 - 0.93	0.45 - 0.61	0.90 - 1.00
	4	0.62	0.41 - 0.80	0.78 - 0.91	0.30 - 0.50	0.86 - 1.10	0.72	0.53 - 0.85	0.86 - 0.94	0.41 - 0.58	0.93 - 1.06
	5	0.67	0.48 - 0.85	0.80 - 0.92	0.33 - 0.55	0.91 - 1.06	0.75	0.60 - 0.89	0.87 - 0.95	0.42 - 0.58	0.95 - 1.04
	6	0.65	0.40 - 0.81	0.78 - 0.91	0.29 - 0.50	0.91 - 1.08	0.77	0.62 - 0.87	0.88 - 0.95	0.44 - 0.60	0.96 - 1.03
	7	0.63	0.42 - 0.81	0.77 - 0.91	0.27 - 0.49	0.90 - 1.09	0.80	0.67 - 0.91	0.89 - 0.96	0.46 - 0.64	0.95 - 1.03
	8	0.64	0.44 - 0.86	0.77 - 0.90	0.26 - 0.48	0.89 - 1.09	0.74	0.56 - 0.90	0.85 - 0.93	0.39 - 0.55	0.93 - 1.05
	9	0.70	0.54 - 0.85	0.83 - 0.94	0.34 - 0.56	0.93 - 1.05	0.86	0.76 - 0.94	0.92 - 0.97	0.56 - 0.73	0.96 - 1.02
	10	0.64	0.38 - 0.84	0.71 - 0.89	0.26 - 0.45	0.86 - 1.09	0.75	0.57 - 0.89	0.82 - 0.92	0.37 - 0.56	0.91 - 1.04
200	1	0.63	0.44 - 0.83	0.73 - 0.86	0.23 - 0.42	0.87 - 1.08	0.77	0.64 - 0.90	0.84 - 0.92	0.37 - 0.55	0.91 - 1.02
	2	0.68	0.47 - 0.87	0.80 - 0.90	0.32 - 0.48	0.90 - 1.05	0.77	0.61 - 0.90	0.87 - 0.93	0.40 - 0.56	0.93 - 1.01
	3	0.62	0.40 - 0.87	0.78 - 0.90	0.25 - 0.45	0.90 - 1.05	0.85	0.73 - 0.97	0.89 - 0.95	0.47 - 0.63	0.94 - 1.01



	4	0.62	0.41 - 0.80	0.77 - 0.90	0.27 - 0.44	0.88 - 1.06	0.72	0.53 - 0.85	0.84 - 0.93	0.36 - 0.52	0.93 - 1.04
	5	0.67	0.48 - 0.85	0.81 - 0.91	0.27 - 0.50	0.92 - 1.04	0.75	0.60 - 0.89	0.86 - 0.93	0.36 - 0.55	0.94 - 1.04
	6	0.65	0.40 - 0.81	0.82 - 0.91	0.32 - 0.49	0.93 - 1.06	0.77	0.62 - 0.87	0.89 - 0.95	0.43 - 0.60	0.95 - 1.03
	7	0.63	0.42 - 0.81	0.79 - 0.90	0.28 - 0.45	0.92 - 1.06	0.80	0.67 - 0.91	0.90 - 0.95	0.46 - 0.62	0.96 - 1.03
	8	0.64	0.44 - 0.86	0.79 - 0.91	0.29 - 0.47	0.90 - 1.06	0.74	0.56 - 0.90	0.86 - 0.94	0.38 - 0.55	0.95 - 1.03
	9	0.70	0.54 - 0.85	0.84 - 0.93	0.36 - 0.52	0.93 - 1.04	0.86	0.76 - 0.94	0.93 - 0.97	0.57 - 0.68	0.97 - 1.01
	10	0.64	0.38 - 0.84	0.79 - 0.90	0.28 - 0.46	0.90 - 1.05	0.75	0.57 - 0.89	0.87 - 0.94	0.40 - 0.56	0.94 - 1.03
	1	0.63	0.46 - 0.78	0.78 - 0.88	0.26 - 0.46	0.89 - 1.04	0.77	0.65 - 0.88	0.87 - 0.93	0.40 - 0.58	0.94 - 1.02
	2	0.68	0.52 - 0.80	0.82 - 0.91	0.29 - 0.49	0.92 - 1.03	0.77	0.58 - 0.89	0.82 - 0.96	0.39 - 0.58	0.82 - 0.96
	3	0.62	0.40 - 0.75	0.78 - 0.89	0.28 - 0.43	0.91 - 1.04	0.85	0.75 - 0.91	0.92 - 0.96	0.54 - 0.66	0.96 - 1.01
	4	0.62	0.46 - 0.76	0.76 - 0.90	0.25 - 0.43	0.90 - 1.08	0.72	0.59 - 0.82	0.84 - 0.93	0.37 - 0.52	0.93 - 1.04
250	5	0.62	0.47 - 0.80	0.79 - 0.89	0.25 - 0.44	0.91 - 1.04	0.75	0.59 - 0.84	0.84 - 0.92	0.34 - 0.51	0.93 - 1.02
	6	0.65	0.51 - 0.81	0.82 - 0.91	0.32 - 0.46	0.93 - 1.04	0.77	0.69 - 0.90	0.87 - 0.93	0.42 - 0.56	0.93 - 1.01
	7	0.63	0.45 - 0.81	0.78 - 0.88	0.26 - 0.45	0.91 - 1.04	0.80	0.60 - 0.91	0.83 - 0.95	0.37 - 0.62	0.83 - 0.95
	8	0.64	0.48 - 0.82	0.80 - 0.90	0.28 - 0.45	0.92 - 1.07	0.74	0.61 - 0.87	0.86 - 0.94	0.37 - 0.53	0.94 - 1.03
	9	0.70	0.55 - 0.81	0.83 - 0.91	0.33 - 0.50	0.92 - 1.02	0.86	0.78 - 0.92	0.92 - 0.96	0.54 - 0.67	0.96 - 1.00
	10	0.64	0.43 - 0.83	0.74 - 0.86	0.21 - 0.41	0.85 - 1.03	0.75	0.61 - 0.89	0.80 - 0.88	0.31 - 0.47	0.88 - 0.99
	1	0.63	0.48 - 0.74	0.78 - 0.88	0.29 - 0.45	0.92 - 1.03	0.77	0.66 - 0.83	0.87 - 0.93	0.41 - 0.55	0.96 - 1.02
	2	0.68	0.51 - 0.79	0.82 - 0.91	0.30 - 0.44	0.90 - 1.02	0.77	0.62 - 0.84	0.88 - 0.94	0.38 - 0.52	0.91 - 1.00
	3	0.62	0.44 - 0.76	0.78 - 0.89	0.26 - 0.41	0.91 - 1.03	0.85	0.74 - 0.91	0.92 - 0.96	0.48 - 0.59	0.95 - 1.01
	4	0.62	0.41 - 0.75	0.76 - 0.90	0.24 - 0.30	0.87 - 1.02	0.72	0.57 - 0.81	0.84 - 0.93	0.32 - 0.44	0.89 - 0.99
350	5	0.67	0.48 - 0.77	0.79 - 0.89	0.30 - 0.46	0.93 - 1.03	0.75	0.62 - 0.83	0.84 - 0.92	0.39 - 0.54	0.95 - 1.02
	6	0.65	0.46 - 0.75	0.82 - 0.91	0.31 - 0.44	0.92 - 1.03	0.77	0.64 - 0.82	0.89 - 0.94	0.41 - 0.55	0.94 - 1.01
	7	0.63	0.42 - 0.73	0.78 - 0.88	0.31 - 0.46	0.93 - 1.03	0.80	0.69 - 0.86	0.87 - 0.93	0.48 - 0.60	0.96 - 1.02
	8	0.64	0.46 - 0.75	0.80 - 0.90	0.30 - 0.45	0.92 - 1.03	0.74	0.58 - 0.83	0.86 - 0.94	0.40 - 0.52	0.95 - 1.02
	9	0.70	0.55 - 0.80	0.83 - 0.91	0.36 - 0.52	0.95 - 1.02	0.86	0.78 - 0.91	0.92 - 0.96	0.57 - 0.67	0.97 - 1.01
	10	0.64	0.47 - 0.75	0.74 - 0.86	0.31 - 0.44	0.92 - 1.04	0.75	0.60 - 0.82	0.80 - 0.88	0.39 - 0.54	0.96 - 1.03

<sup>a</sup> Confirmatory factor analysis

<sup>b</sup> Number of observations



- <sup>c</sup> True value of loadings
- <sup>d</sup> Model based estimated loadings
- <sup>e</sup> Single selected day
- <sup>f</sup> Weekly item average
- <sup>g</sup> Within-individual analysis (the estimates are based on multilevel CFA)
- <sup>h</sup> Between-individual analysis (the estimates are based on multilevel CFA)



### 6.2.2 Estimated, true and observed inter-item correlation

Estimated correlation matrices as determined via the iterative CFAs across the simulated datasets for all scenarios and various sample sizes, and they were compared with the true correlation matrix for the weekly item average and single selected day approaches (see Table 6.6 and see Appendix A for the other factors). Comparing this with the EFA results, we can see a similar pattern as before, with the item average approach tending towards a narrower range and generating higher estimates compared to the single day approach and the true range of the correlation matrix parameters. This was also well reflected in Figure 6.10 where the distribution of the range of the estimated, true and observed correlation is visualized (see Appendix B for the other factors).



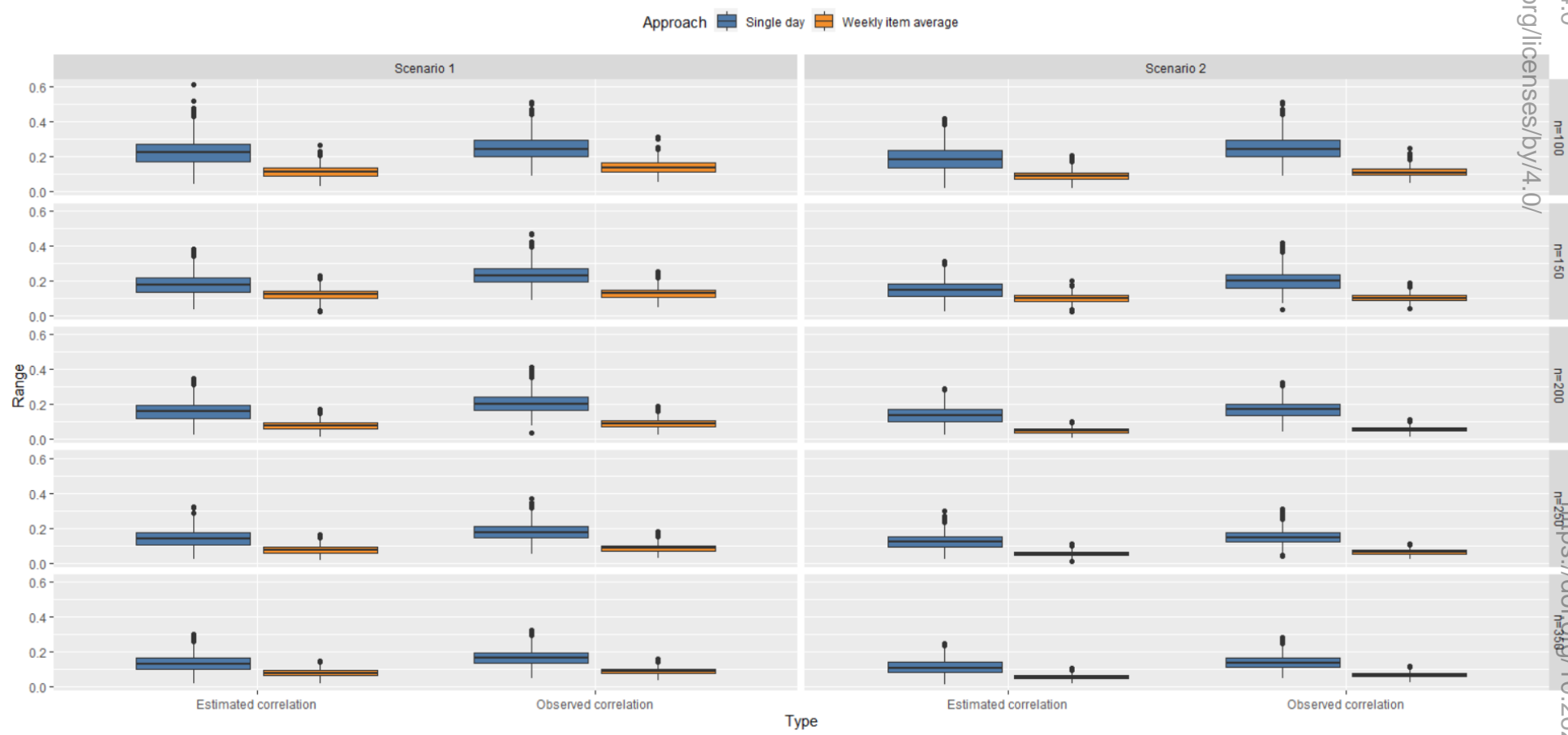
Table 6.6: True correlation parameters, observed inter-item correlations, and CFA-estimated correlations for the single selected day and weekly item average approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

n <sup>b</sup>	Factor	Scenario 1					Scenario 1					
		$R_T^c$		$R_M^d$		$R_O^e$		$R_T^c$		$R_O^e$		
		SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>	SD <sup>f</sup>	WIA <sup>g</sup>	
100	1	0.39 - 0.48	0.57 - 0.86	0.11 - 0.71	0.53 - 0.84	-0.08 - 0.78	0.52 - 0.88	0.55 - 0.74	0.26 - 0.84	0.70 - 0.93	0.10 - 0.84	0.67 - 0.94
	2	0.41 - 0.49	0.39 - 0.83	0.01 - 0.72	0.58 - 0.84	-0.24 - 1.00	0.30 - 0.83	0.51 - 0.71	0.12 - 0.87	0.55 - 0.89	-0.09 - 1.00	0.45 - 0.89
	3	0.39 - 0.49	0.59 - 0.87	0.11 - 0.83	0.65 - 0.85	-0.02 - 0.81	0.51 - 0.88	0.53 - 0.60	0.26 - 0.87	0.69 - 0.92	0.23 - 0.88	0.63 - 0.92
150	1	0.39 - 0.48	0.53 - 0.84	0.19 - 0.67	0.53 - 0.84	0.07 - 0.74	0.49 - 0.86	0.55 - 0.74	0.36 - 0.86	0.72 - 0.91	0.27 - 0.87	0.66 - 0.92
	2	0.41 - 0.49	0.58 - 0.84	0.18 - 0.75	0.58 - 0.84	0.11 - 0.74	0.54 - 0.84	0.51 - 0.71	0.31 - 0.87	0.63 - 0.91	0.15 - 0.87	0.60 - 0.92
	3	0.39 - 0.49	0.59 - 0.87	0.23 - 0.78	0.65 - 0.85	0.13 - 0.79	0.57 - 0.88	0.53 - 0.60	0.37 - 0.82	0.69 - 0.90	0.29 - 0.84	0.69 - 0.90
200	1	0.39 - 0.48	0.58 - 0.83	0.21 - 0.66	0.58 - 0.83	0.09 - 0.71	0.57 - 0.84	0.55 - 0.74	0.39 - 0.82	0.73 - 0.92	0.29 - 0.84	0.70 - 0.93
	2	0.41 - 0.49	0.60 - 0.82	0.18 - 0.67	0.60 - 0.82	0.10 - 0.65	0.57 - 0.83	0.51 - 0.71	0.33 - 0.82	0.70 - 0.90	0.21 - 0.82	0.69 - 0.91
	3	0.39 - 0.49	0.65 - 0.85	0.27 - 0.73	0.65 - 0.85	0.18 - 0.70	0.65 - 0.85	0.53 - 0.60	0.38 - 0.79	0.75 - 0.88	0.37 - 0.79	0.74 - 0.88
250	1	0.39 - 0.48	0.59 - 0.81	0.24 - 0.62	0.59 - 0.81	0.11 - 0.68	0.53 - 0.83	0.55 - 0.74	0.39 - 0.80	0.68 - 0.91	0.31 - 1.00	0.63 - 0.93
	2	0.41 - 0.49	0.54 - 0.82	0.22 - 0.65	0.54 - 0.82	0.13 - 0.66	0.51 - 0.82	0.51 - 0.71	0.33 - 0.80	0.65 - 0.89	0.24 - 0.79	0.60 - 0.90
	3	0.39 - 0.49	0.64 - 0.84	0.22 - 0.65	0.64 - 0.84	0.20 - 0.67	0.61 - 0.84	0.53 - 0.60	0.36 - 0.74	0.75 - 0.88	0.36 - 0.75	0.74 - 0.88
350	1	0.39 - 0.48	0.58 - 0.82	0.20 - 0.59	0.58 - 0.82	0.15 - 0.60	0.56 - 0.83	0.55 - 0.74	0.39 - 0.78	0.69 - 0.90	0.34 - 0.83	0.66 - 0.91
	2	0.41 - 0.49	0.56 - 0.80	0.24 - 0.67	0.56 - 0.80	0.20 - 0.64	0.55 - 0.80	0.51 - 0.71	0.36 - 0.83	0.67 - 0.88	0.31 - 0.81	0.65 - 0.89
	3	0.39 - 0.49	0.65 - 0.82	0.26 - 0.63	0.65 - 0.82	0.20 - 0.63	0.64 - 0.83	0.53 - 0.60	0.39 - 0.73	0.75 - 0.87	0.37 - 0.72	0.73 - 0.87

<sup>a</sup> Confirmatory factor analysis  
<sup>b</sup> Number of individuals  
<sup>c</sup> Range of the true correlation between items  
<sup>d</sup> Range of the model-based estimated correlation between items under the EFA model  
<sup>e</sup> Range of the observed correlation between items  
<sup>f</sup> Single selected day  
<sup>g</sup> Weekly item average



Figure 6.10: Boxplot of the range of inter item observed and CFA-estimated correlation within the first factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day approach; Item average=Weekly item average; a: Slope parameter from a multidimensional graded response model.



### 6.2.3 Overall factor loading bias and mean squared error

With the simulated data, a similar pattern of bias was observed across the weekly item average and single day-based implementations of CFA as was seen in the prior EFA stage of this work (see Figure 6.11). The weekly item average approach showed higher overall bias and MSE higher compared to the single selected day approach. The prior investigation of the range of the factor loadings revealed that weekly item average was prone to overestimation in contrast with the single selected day which did not show any systematic pattern due to the wide range of the estimated loadings. Additionally, the MSE (see Figure 6.12) showed a similar pattern with EFA, with the weekly item average producing higher mean square error compared to the single selected day.



Figure 6.11: Overall absolute average bias of the estimated loadings of items within the first factor for the CFA model with 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory Factor Analysis; Single day=Single selected day approach; Item average=Weekly item average; a: Slope parameter from a multidimensional graded response model.

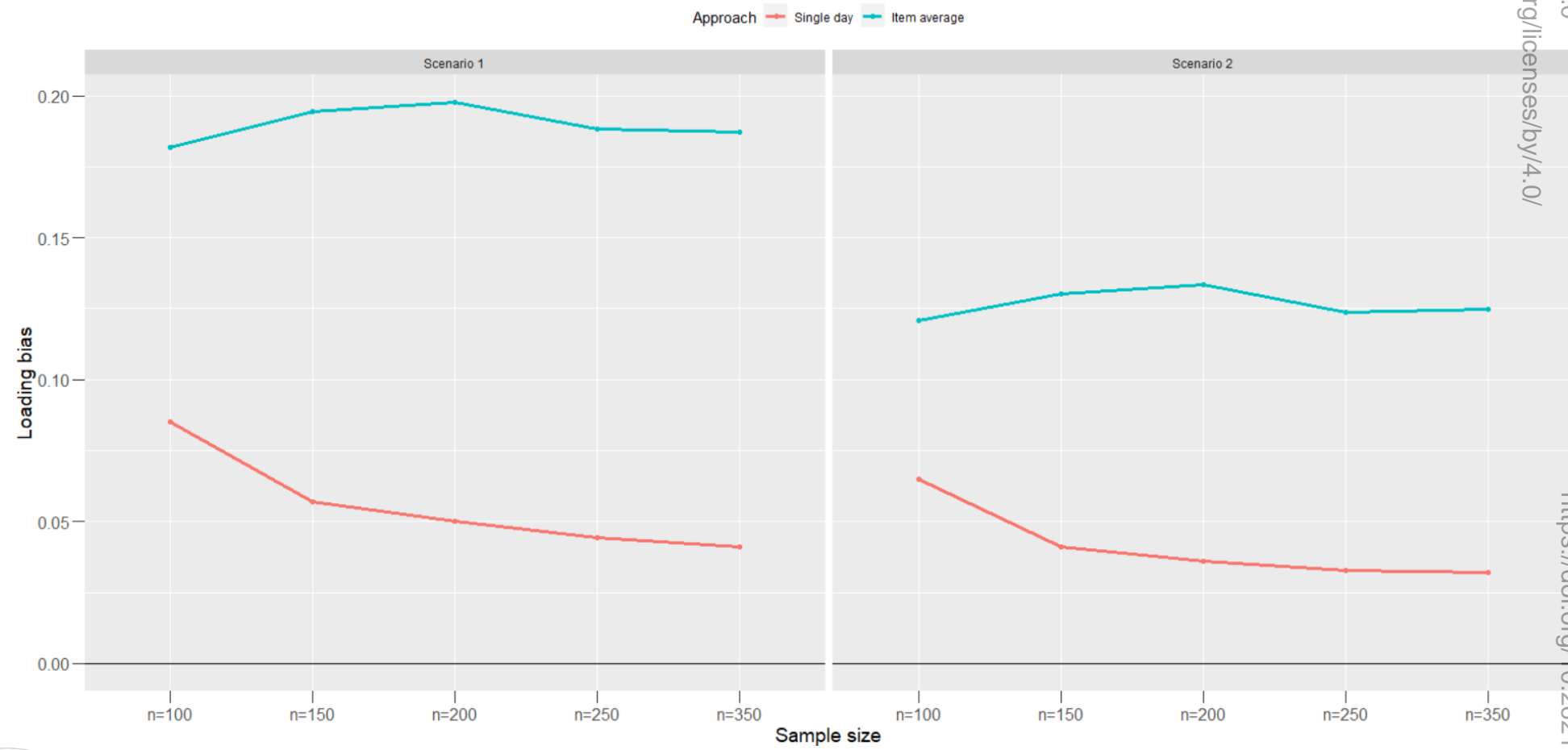
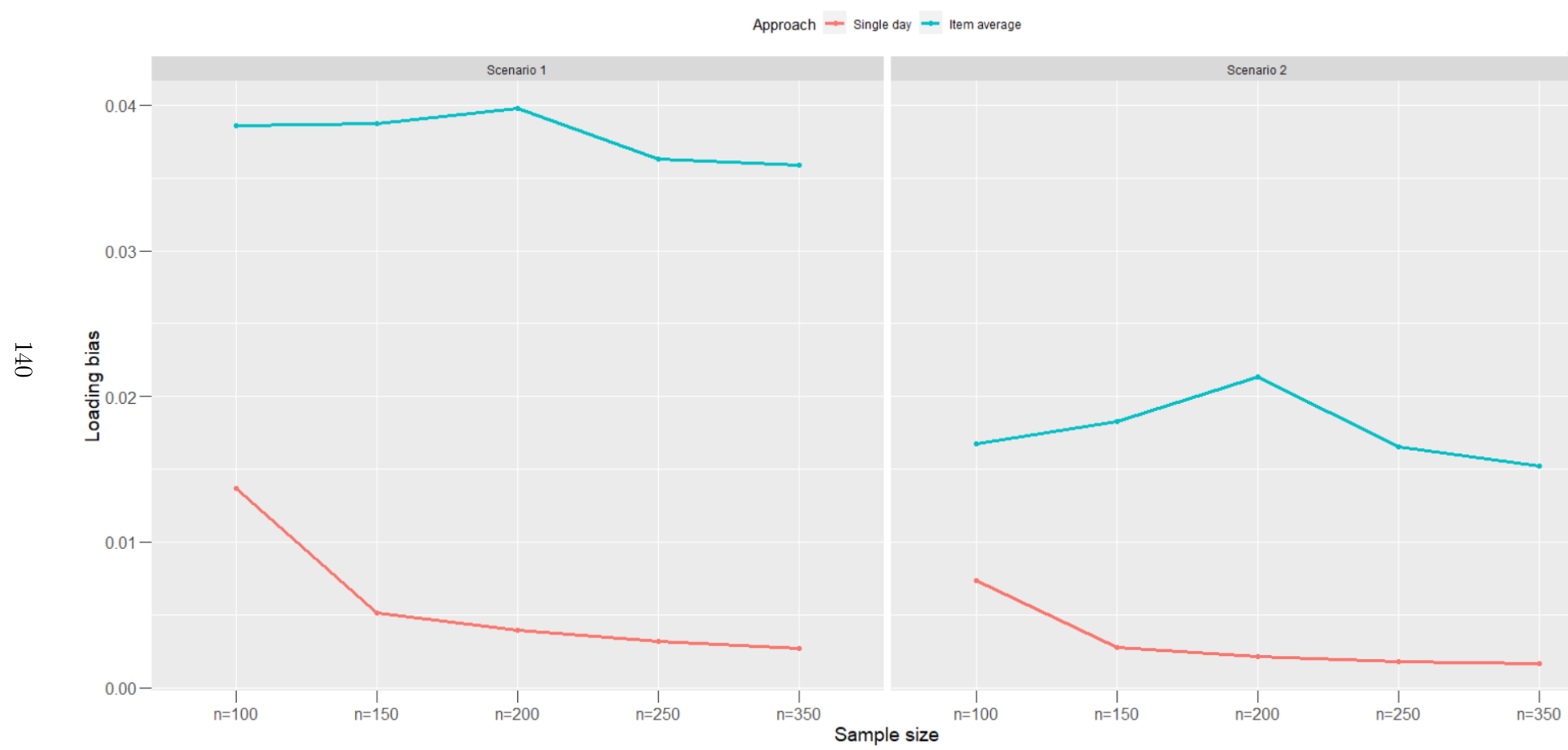


Figure 6.12: Overall MSE of the estimated loadings of items of within the first factor for the CFA model with 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day approach; Item average=Weekly item average; a: slope parameter.



## 6.2.4 Goodness of fit measures

The evaluation of the goodness-of-fit measures with the CFA implementations (see 6.13, 6.14 and 6.15) yielded similar conclusions to what had been observed with the earlier EFAs, but there were also some difference. One of the common pattern across both EFA and CFA was the behaviour of SRMR which showed more frequently an acceptable fit than a good fit for the single selected day approach. However, one notable difference was the fact that the single selected day approach showed the highest occurrence of bad fit in comparison with the other 2 approaches based on CFI, TLI and SRMR on both scenarios, with the percentages being decreased more fast for scenario 2 as the sample size increased. RMSEA also showed a high percentage of acceptable fit for the weekly item average approach in comparison with the other fit measures.

A notable difference of the CFA evaluation compared to EFA evaluation was that the multilevel CFA consisted of an overall fit of the model, rather than an evaluation of the within- and between-individual model separately. Figure 6.15 demonstrated that the multilevel CFA almost universally achieved a good fit based on all goodness of fit measures across both scenarios. Although the occurrence of good fit was expected as such a model could capture the multilevel nature of the data, there was also a serious consideration regarding the overfitting of the model. Even though they were many datasets, some of which produced quite weak factor structure as suggested on the prior analysis, the multilevel managed to reproduce the correlation matrix perfectly as the result of overparametrization of the model.

A final inspection on the difference between the two scenarios showed that CFI and TLI had lower percentages of good fit for scenario 1 compared to scenario 2 for the single selected day. Those measures necessitated a greater sample size to produce higher frequencies of good fit for scenario 1 as the results of the weaker loadings compared to scenario 2.

Overall, the results demonstrated that divergent insights could be retrieved from goodness of fit measures that assess different aspects of the model. Some highlights of the results were the high percentage of RMSEA for the single selected day, the high percentage of good fit for CFI for the weekly item average approach, the high percentage of acceptable fit for the single selected day based on SRMR, and the very high percentage of good fit on the multilevel model, a part of which may be attributed to its overfitting.



Figure 6.13: Bar plots for the percentage of poor fit (see Chapter 4.8) in CFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model.



Figure 6.14: Bar plots for the percentage of acceptable fit (see Chapter 4.8) in CFA of the 1,000 simulated datasets for each of the data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model.

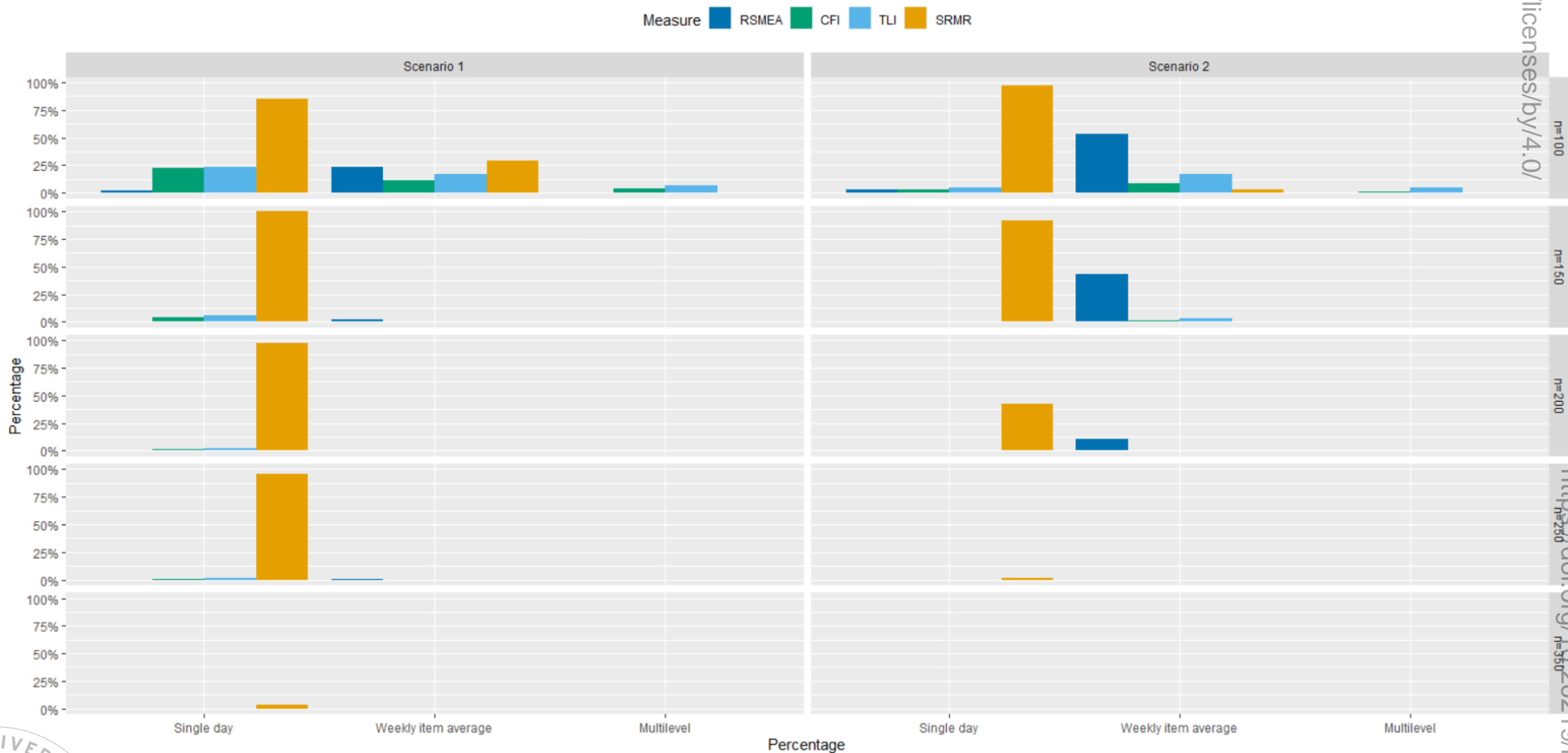
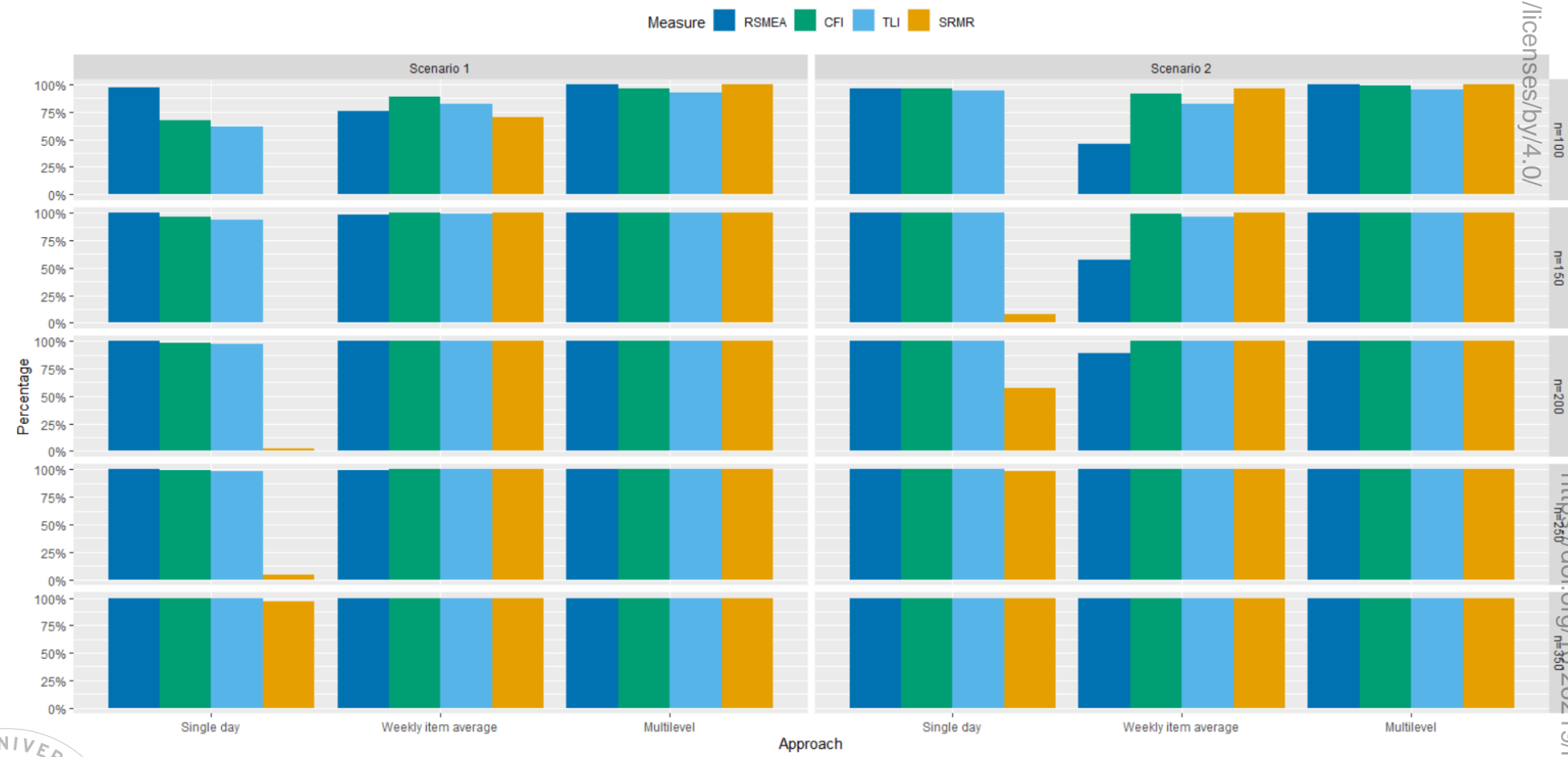


Figure 6.15: Bar plots for the percentage of good fit (see Chapter 4.8) in CFA of the 1,000 simulated datasets for each of the data handling approaches for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ) with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; RMSEA=Root mean square error of approximation; CFI=Comparative fit index; TLI=Tucker-Lewis index; SRMR=Standardized root mean squared residual; a: Slope parameter from a multidimensional graded response model.



## 6.2.5 Convergence and Heywood cases

The final focus of this analysis was dedicated to the occurrence of Heywood cases and convergence issues across the 3 approaches. A notable difference with the prior EFA analysis was the different estimation process of the multilevel CFA where both between- and within-individual model were estimated simultaneously, which was more computationally demanding than the split EFA approach. The convergence of multilevel CFA ranged from 88.1% to 94.9% for scenario 1 while for scenario 2 it ranged from 91.3% to 93.9%. Also, for  $n=100$  the single selected day approach seemed to have some convergence issues across both scenarios (i.e., 99.2% for scenario 1 and 99.9% for scenario 2). Another apparent issue for multilevel CFA was Heywood case, which remained on high levels across both scenarios. For scenario 1 it ranged from 81.2% to 93.7% and for scenario 2 it ranged from 76.4% to 91.5%. The reason for its occurrence was attributed to the between-individual model, which utilized a lower sample size relative to the within-individual model. The dependency of Heywood case with the number of individuals was well reflected in Table 6.7 as the increase of the sample size on the between-individual level resulted in the reduction of Heywood cases. Such an occurrence is a serious issue as it could lead to improper solutions, thus the interpretation of the results could be misleading.

Overall, the results demonstrated that multilevel CFA entails important methodological issues when it comes to convergence issues and Heywood cases compared to the other simplistic approaches.



Table 6.7: Percentage of convergence and Heywood case for the single selected day, weekly item average, within- and between-individual analysis approaches for CFA<sup>c</sup>model with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

	n	Scenario 1			Scenario 2		
		SD <sup>a</sup>	WIA <sup>b</sup>	Multilevel	SD <sup>a</sup>	WIA <sup>b</sup>	Multilevel
Convergence	100	99.2%	100.0%	93%	100.0%	99.9%	91.3%
	150	100.0%	100.0%	88.1%	100.0%	100.0%	86.2%
	200	100.0%	100.0%	93.0%	100.0%	100.0%	89.5%
	250	100.0%	100.0%	94.9%	100.0%	100.0%	89.5%
	350	100.0%	100.0%	90.9%	100.0%	100.0%	93.9%
Heywood case	100	0.7%	0.0%	93.3%	0.1%	0.0%	91.5%
	150	0.0%	0.0%	87.6%	0.2%	0.0%	85.2%
	200	0.0%	0.0%	93.7%	0.0%	0.0%	89.2%
	250	0.0%	0.0%	86.8%	0.0%	0.0%	76.4%
	350	0.0%	0.0%	81.2%	0.0%	0.0%	81.2%

<sup>a</sup> Single selected day

<sup>b</sup> Weekly item average

<sup>c</sup> Confirmatory factor analysis

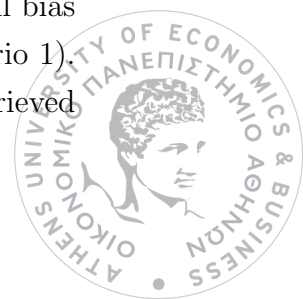


# Chapter 7

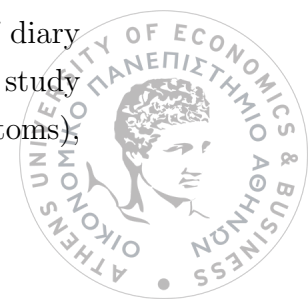
## Discussion and Conclusions

### 7.1 Discussion on the Results

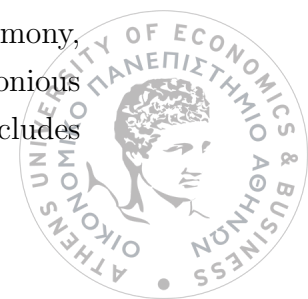
Through comparative evaluation of different strategies for factor analysing daily diary data using simulated data both across two different scenarios (i.e. scenario 1: the true loading parameters were moderate; scenario 2: the true loading parameters were strong) and different sample sizes, the present study demonstrated the inferential implications of each modelling approach. This simulation study showed divergent findings and evidence of potential methodological issues in the results across all the data handling approaches, which highlights the need for more careful consideration of the approaches used when proceeding to the structural evaluation of a daily diary instrument. When the focus is confined to a single day under an exploratory modelling framework, the results highlight the risk for incorrect identification of the number of factors, particularly when assessed via the widely used Kaiser criterion, whereas parallel analysis showed a quite good performance. When comparing Kaiser criterion and empirical Kaiser criterion, the latter demonstrated a better performance within the context of small sample sizes (e.g.,  $n=100$ ,  $n=150$ ,  $n=200$ ). Our findings on the weekly item average approach suggested that it may be prone to an overestimation of factor loadings as the result of the increased observed correlation coefficients under both EFA and CFA frameworks. The results also showed that for the weekly item average in scenario 2 where the true loading strength was quite high there were occurrences of Heywood cases (i.e., standardized loadings greater than 1) under small sample sizes (i.e.,  $n=100$ ,  $n=150$ ,  $n=200$ ). Evidence was also established that the overall absolute average bias and MSE was higher for the weekly item average approach compared to the single selected day approach. The former approach demonstrated that the overall bias and MSE of might be higher when the true loading strength is moderate (scenario 1). Our findings for multilevel models showed that additional insights could be retrieved



for the evaluation of daily diary data, such as assessing the property of dimensional invariance. Such a property is usually not checked in applied research, although the construct across the two levels of the instrument might not always have the same interpretation. In addition to the potential benefits of multilevel models, there were also some methodological issues and limitations. More specifically, under the CFA framework, there was a high occurrence of Heywood cases and convergences compared to the single-level approaches. An additional consideration was also the increased risk of overfitting as a result of the overparameterized model. An interesting result in the evaluation of the simulated data across both scenarios was the contradiction in evidence of fit offered by the SRMR and other goodness of fit measures such as RMSEA and TLI. For example, in the simulated data, the SRMR would suggest an acceptable or good fit even in cases where the estimated and true correlation matrices were notably divergent (as in the case of the weekly item average approach). By contrast, in such instances, RMSEA and TLI would be more conservative, which was expected based on previous studies (Hu & Bentler, 1999). Consequently, SRMR should not be used as a stand-alone goodness of fit measure, but it should be used with other measures as well such as RMSEA, TLI and CFI. The divergent findings and evidence of all approaches (and limitations across each of the data handling approaches) underlines why careful consideration should be given to the strategy adopted in structural evaluation of daily diary instruments. The application of each method should be informed by the intended research context and aims. There are instances, for example, where the use of a derived variable approach (e.g., averaging observations across time) or selecting a random day can sidestep the challenges inherent in implementing modelling strategies for evaluating the latent structure of intensive longitudinal data. The selected single day approach seems an adequate and appropriate method when the main goal is to understand the relationship between-individual items and where longitudinal measurement invariance can be assumed (such that within-individual variability can be validly considered as a statistical nuisance). The item weekly average approach on the other hand may be applicable to study the patterns of association between items as aggregated over time. For example, this approach could be adopted where the instrument being evaluated is intended for scoring at an aggregate level and interest is principally on studying of how individuals differ on average across a study period. Additionally, the multilevel approach is recommended to be used supplementary with the other approaches for assessing additional psychometric properties of an instrument such as dimensional and cross-level invariance. There are also practical and contextual considerations when it comes to selecting such data handling strategies for the structural evaluation of diary data. For example, in a PRO assessment program, where the main purpose of a study is to monitor symptoms that are intermittently experienced (e.g., asthma symptoms),



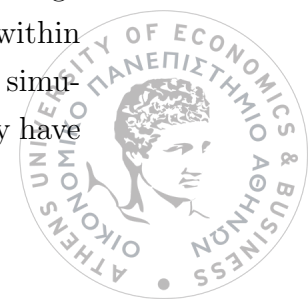
a decision is required whether the average score should be derived only on occasions when the symptom is present or at all measured timepoints. These two approaches could lead to different results and conclusions (Stone et al., 2012) although such event- vs time-based considerations were not examined within the context of the present study. Importantly, when within-individual variance is not considered a statistical nuisance but rather relevant to the research question at hand, then both the weekly item average and single selected day strategies ignore or mask such pertinent detail. This is an important consideration, as both within- and between-individual variances are important to understanding how rates of change over time may differ between individuals or groups. By decomposing variance into both a within- and between-individual component, multilevel models, enable the examination intra- and inter-individual of dimensional invariance. Such properties can inform the presence or absence of ecological fallacy. These assumptions are often untested in cross-sectional applications of factor analysis, risking erroneous inference. An examination of ICC coefficients served to establish that both sources of variance were indeed present in the simulated datasets (as per their design). This illustrated the importance of adopting multilevel modelling approaches that are designed to account for and provide insight into variance at both levels. Subsequent implementation of multilevel approaches to factor analysis in the present study highlighted the additional insights that could be obtained via this set of techniques when it comes to understanding latent structures in diary data under both exploratory and confirmatory frameworks. A consideration when conducting exploratory multilevel factor analysis based on the split modelling approach used in the present study is the fact that the sample between-individual covariance matrix is not an unbiased estimation of the between population covariance matrix. As a consequence, the structural model as established at the sample level might not always be an accurate reflection of the structure of the population between-individual covariance matrix. Although there is an unbiased estimation approach, its adoption remains uncommon as it usually results in a non-positive definite matrix. Another consideration when conducting multilevel models is that they do not account for the day-to-day correlation in the factor scores. Dynamic factor models (P. Molenaar, 1985), an extension of the multilevel modelling approaches, could potentially be used as an alternative strategy. Given that they account for additional complexity such as latent correlation patterns, they may provide further useful insights (e.g. how the score of the first factor at time  $t$  affects the score of the second factor at time  $t+1$ ). This approach was, however, beyond the scope of the present study. An additional consideration when conducting multilevel factor analysis is the trade-off between goodness of fit measures and model parsimony, as a less parsimonious model will tend to have a better fit than a more parsimonious model. This is directly related to the issue of overfitting as the multilevel model includes



a high volume of parameters, including within-, between-level loadings and within- and between-level error variances. That is a very crucial issue when evaluating the factor structure of the data, as this could result in misleading results as the results may not be replicable in different dataset. This is because when a model is overfitted, it tends to perfectly fit the current data under analysis, which poses an obstacle to the generalization of the results. An additional issue, which is not often discussed, is that factor models are generally subject to rotational indeterminacy. This means that factors and loadings are identified only up to a rotation. As a result there will be infinitely possible solutions (e.g., estimated loadings), which will be all equivalent. Such an issue is more pertinent in the EFA framework, as in CFA framework the model is subject to some restrictions (e.g., some loadings are fixed to zero or one). Conversely in EFA approach loadings needs to be estimated for each factor and item. A solution to this problem was proposed by Thurstone (1954), who introduced the simple structure, which provides a set of rules that need to hold in order the solution to be acceptable and yield interpretable results. The more realistic these rules are compared to the actual truth, the more accurate results can be obtained. In fact this proposed method have been shown to yield accurate results (e.g., (Sokal, 1958)). This is the reason for which this solution was also proposed for this simulation study by using oblimin rotation. Even so, the selection of this rotation may still seem somewhat arbitrary (there are limited studies providing a comparative evaluation of the different rotation methods (Finch, 2011)) and in some scenarios where the items are related with multiple factors (unlike this simulation study), such a solution may be inadequate and provide misleading results. For example, for more complex loading patterns it has been shown that when using the common rotation methods, the cross-loadings are underestimated and factor correlations are inflated (Scharf & Nestler, 2019). Overall, the results demonstrate that a multilevel strategy could provide additional insight when investigating the latent structure of diary measures. However, there are still important methodological issues that impact the proper interpretation of the results, posing adherence challenges to regulatory requirements. The first two notable issues were the occurrence of Heywood cases and convergence failures. Heywood cases result in standardized loadings that appear greater than 1 (Kolenikov & Bollen, 2012). This generally could be attributed to multiple reasons such as outliers (Bollen, 1987), underidentification (Boomsma & Hoogland, 2001), structurally misspecified models (Dillon, Kumar, & Mulani, 1987) or sampling fluctuations (Van Driel, 1978). In the present study, this may have been attributable to sample error resulting from the sample size parameter used in the simulation which could mean that the population variance is positive and near zero and the corresponding negative estimates of the variances are attributed to chance (Kolenikov & Bollen, 2012) as the risk of Heywood case occurrence is sample size dependent (Bartholomew et al.



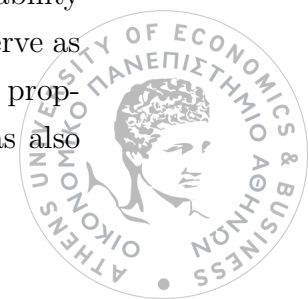
2011). Generally, there have been a lot of proposed solutions, which are well summarized here (Farooq, 2022). One of them is to set the range of the error variance estimates to  $[0, +\infty)$ . Although such a method is simple, there are many issues to be considered as has been discussed (Kolenikov & Bollen, 2012), as the MLE estimator do not share good properties near 0, which leads to unstable estimators. It is evident that Heywood and convergence issues showed higher percentage of occurrence on multilevel CFA in comparison with split approach. This was ascribed to the estimation process of the within- and between-individual models. The split approach estimates the between- and within-individual model separately (B. O. Muthén, 1994) and the multilevel approach estimates the between- and within-individual model simultaneously, which accounts for the fact that the sample between-individual covariance matrix is an unbiased estimator of the linear combination of the population within and between-individual model (J. Hox & Maas, 2004). Although split approach is not appropriate under CFA framework, it could still be a useful strategy to explore and get an initial view of the factor structure of the data with less Heywood cases and convergence issues than multilevel approach. Another issue that required a further consideration, especially in the context of the regulatory demands, is the dependency of the within- and between- individual models. Even in cases where the within-individual model is correctly specified, any misspecification of the between-individual model could significantly impact the within-level estimates and its fit, which could raise serious concerns for the appropriateness of both latent variable models. This issue warrant further attention given that the specification of the between-individual model is quite challenging, as has been previously flagged. Such issues pose great challenges when assessing and validating a diary PRO instrument, especially when interacting with regulators, as most of the time one should define a priori the scoring algorithm of an instrument. However, the structural validation evidence for most diary PRO instruments is typically based on evidence at a single level, and the factor structures at both the within- and between-individual level is therefore usually not known a priori. There are additional considerations and limitations to the present study that necessitate acknowledgement. First, as mentioned, the data simulation was implemented under 5 different selected small sample sizes under 2 scenarios, with the intention being to reflect the clinical dataset sizes often encountered in early phase health research. As a consequence, however, the findings might have limited generalizability to contexts where larger sample sizes are more feasible. Secondly, estimation methods for explanatory and confirmatory multilevel factor analysis of the simulated data were implemented under the assumption of continuous data, as to our knowledge there is currently no package for conducting ordinal multilevel factor analysis within the analytic software used in this study (R Core Team et al., 2013). Thirdly, the simulated data were based on a 5-point response scaling, meaning this assumption may have



an impact on the goodness of fit statistics and the standard errors of the parameter estimates (Chou & Bentler, 1995). This also means that the conventional cut-off values that were assumed to be the same for all the approaches might not be the optimum strategy. It is possible that for the single day approach (where the data being modelled were on an ordinal scale), a model with a poor fit may be accepted more frequently in comparison with the weekly item average approach, where the data are continuous (Xia & Yang, 2019). Along with these cut-off values for the goodness of fit measures, the threshold to assess whether an item loads on its corresponding factor is another limitation of this study in the sense that alternative options (Field, 2005; Stevens, 2012) were not evaluated. The suitability of different fit indices to categorical data remains an area of discussion and debate (see for example alternative perspectives on the suitability of SRMR to fit ascertainment with such data: (Brown, 2015), (Shi et al., 2020), (Yu, 2002)). As a fourth limitation, due to the differing scales/measurement levels, the input matrices in the simulation study differed between the single day approach (where polychoric correlations were used) and for the weekly item average (where Pearson correlations were used). This may influence differences between the approaches in the subsequent findings. Fifth, in the simulated data analyses, the same dataset was used in both EFA and CFA, preventing the use of independent subsets for cross-validation of findings. However, the main focus of the present study was not on development or validation of an instrument and given the limited sample sizes explored, such dataset splitting was not considered pivotal to the aims or practicalities of the present evaluation. Finally, missingness was not taken into account in the current study, so the results between the different approaches could differ under its presence.

## 7.2 Concluding Remarks

Through an in-depth and wide-ranging evaluation, this study has illustrated the important empirical and inferential consequences of different factor analytic strategies as applied to daily diary data. Diary data are an important modality for PRO assessment as they offer researchers invaluable insights regarding the association patterns and increasing/decreasing trends of patient's symptoms. The importance of such data and how to properly study their factor structure can become more pronounce when considering the possible consequences and diverse insights that can be obtained through different data handling strategies under real-world data. Methods that only model between-individual variability can offer relevant insight, but do not account for within-person variability that may be comparably informative and pertinent. Multilevel approaches can serve as an alternative and effective strategy for simultaneously evaluating measurement properties at both the between- and within-individual level. However, our study has also

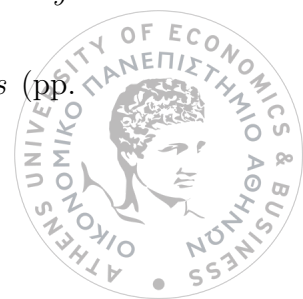


shown that such multilevel modelling entails methodological challenges which require recognition and consideration. The learning established in this work may also generalize to other longitudinally comparative research needs, such as other psychometric property assessment (reliability, validity, and responsiveness) and intensive longitudinal data-based efficacy analysis.



## References

- Anker, S. D., Agewall, S., Borggreffe, M., Calvert, M., Jaime Caro, J., Cowie, M. R., ... others (2014). The importance of patient-reported outcomes: a call for their comprehensive integration in cardiovascular clinical trials. *European Heart Journal*, *35*(30), 2001–2009.
- Asparouhov, T., & Muthén, B. (2010). Simple second order chi-square correction. *Mplus technical appendix*, 1–8.
- Babakus, E., Ferguson Jr, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, *24*(2), 222–228.
- Baker, F. B. (2001). *The basics of item response theory*. ERIC.
- Baldwin, M., Spong, A., Doward, L., & Gnanasakthy, A. (2011). Patient-reported outcomes, patient-reported information. *The Patient: Patient-Centered Outcomes Research*, *4*(1), 11–17.
- Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.
- Benjamin, K., Vernon, M. K., Patrick, D. L., Perfetto, E., Nestler-Parr, S., & Burke, L. (2017). Patient-reported outcome and observer-reported outcome assessment in rare disease clinical trials: an ispor coa emerging good practices task force report. *Value in Health*, *20*(7), 838–855.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238.
- Bentler, P. M. (2010). Sem with simplicity and accuracy. *Journal of Consumer Psychology*, *20*(2), 215–220.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588.
- Birnbaum, A., Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. *Some latent trait models and their use in inferring an examinee's ability*. Addison-Wesley, Reading, MA.
- Bishop, C. M. (1998). Latent variable models. In *Learning in graphical models* (pp. 371–403). Springer.



- Black, N. (2013). Patient reported outcome measures could help transform healthcare. *British Medical Journal*, 346.
- Bolger, N., Davis, A., & Rafaeli, E. (2003). Diary methods: Capturing life as it is lived. *Annual Review of Psychology*, 54(1), 579–616.
- Bollen, K. A. (1987). Outliers and improper solutions: A confirmatory factor analysis example. *Sociological Methods & Research*, 15(4), 375–384.
- Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.
- Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models* (Vol. 154). Sage.
- Boomsma, A. (2000). Reporting analyses of covariance structures. *Structural Equation Modeling*, 7(3), 461–483.
- Boomsma, A., & Hoogland, J. J. (2001). The robustness of lisrel modeling revisited. *Structural Equation Models: Present and future. A Festschrift in honor of Karl Jöreskog*, 2(3), 139–168.
- Braeken, J., & Van Assen, M. A. (2017). An empirical kaiser criterion. *Psychological Methods*, 22(3), 450.
- Broderick, J. E., Schwartz, J. E., Schneider, S., & Stone, A. A. (2009). Can end-of-day reports replace momentary assessment of pain and fatigue? *The Journal of Pain*, 10(3), 274–281.
- Brose, A., & Ram, N. (2012). Within-person factor analysis: Modeling how the individual fluctuates and changes across time. *American Psychological Association*.
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.
- Browne, M. W. (1968). A comparison of factor analytic techniques. *Psychometrika*, 33(3), 267–334.
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258.
- Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104(3), 396.
- Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4), 509–540.
- Bushnell, D. M., McCarrier, K. P., Bush, E. N., Abraham, L., Jamieson, C., McDougall, F., ... others (2019). Symptoms of major depressive disorder scale: performance of a novel patient-reported symptom measure. *Value in Health*, 22(8), 906–915.
- Byrne, B. M. (2013). *Structural equation modeling with lisrel, prelis, and simplis: Basic concepts, applications, and programming*. Psychology Press.

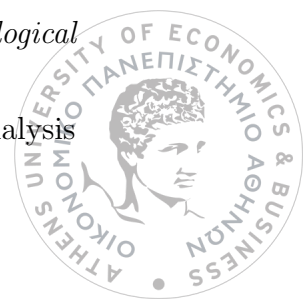


- Carroll, J. B. (1953). An analytical solution for approximating simple structure in factor analysis. *Psychometrika*, *18*(1), 23–38.
- Cattell, R. B. (1963). The structuring of change by p-technique and incremental r-technique. *Problems in Measuring Change*, 167–198.
- Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276.
- Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and kg criteria for determining the number of factors. *Multivariate Behavioral Research*, *12*(3), 289–325.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the r environment. *Journal of Statistical Software*, *48*, 1–29.
- Chen, F., Bollen, K. A., Paxton, P., Curran, P. J., & Kirby, J. B. (2001). Improper solutions in structural equation models: Causes, consequences, and strategies. *Sociological Methods & Research*, *29*(4), 468–508.
- Chou, C.-P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. *American Psychological Association*.
- Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin*, *103*(2), 276.
- Coenders, G., Satorra, A., & Saris, W. E. (1997). Alternative approaches to structural modeling of ordinal data: A monte carlo study. *Structural Equation Modeling*, *4*(4), 261–282.
- Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., . . . Cohen, I. (2009). Pearson correlation coefficient. *Noise Reduction in Speech Processing*, 1–4.
- Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., & Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably? *Personality and Social Psychology Bulletin*, *32*(7), 917–929.
- Cronbach, L. J., & Webb, N. (1975). Between-class and within-class effects in a reported aptitude\* treatment interaction: Reanalysis of a study by gl anderson. *American Psychological Association*.
- Crowley, S. L., & Fan, X. (1997). Structural equation modeling: Basic concepts and applications in personality assessment research. *Journal of Personality Assessment*, *68*(3), 508–531.
- de Haan-Rietdijk, S., Voelkle, M. C., Keijsers, L., & Hamaker, E. L. (2017). Discrete-vs. continuous-time modeling of unequally spaced experience sampling method data. *Frontiers in Psychology*, *8*, 1849.
- Depaoli, S., Tiemensma, J., & Felt, J. M. (2018). Assessment of health surveys: Fitting a multidimensional graded response model. *Psychology, Health & Medicine*,

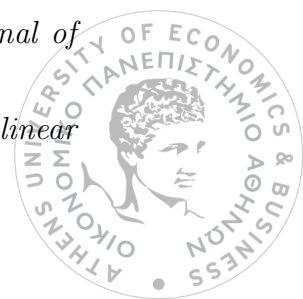


23(sup1), 1299–1317.

- Diamantopoulos, A., & Siguaw, J. (2000). *Introducing lisrel*. london: Sage publications.
- Diggle, P., Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S., et al. (2002). *Analysis of longitudinal data*. Oxford university press.
- Dillon, W. R., Kumar, A., & Mulani, N. (1987). Offending estimates in covariance structure analysis: Comments on the causes of and solutions to heywood cases. *Psychological Bulletin*, 101(1), 126.
- Dinno, A. (2009). Exploring the sensitivity of horn's parallel analysis to the distributional form of random data. *Multivariate Behavioral Research*, 44(3), 362–388.
- Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5 and 7 response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology*, 47(2), 309–326.
- Efron, B., & LePage, R. (1992). *Introduction to bootstrap*. Wiley & Sons, New York.
- Fairclough, D. L., Peterson, H. F., & Chang, V. (1998). Why are missing quality of life data a problem in clinical trials of cancer therapy? *Statistics in Medicine*, 17(5-7), 667–677.
- Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, 6(1), 56–83.
- Farooq, R. (2022). Heywood cases: possible causes and solutions. *International Journal of Data Analysis Techniques and Strategies*, 14(1), 79–88.
- Fayers, P. M., & Machin, D. (2013). *Quality of life: the assessment, analysis and interpretation of patient-reported outcomes*. John Wiley & Sons.
- FDA. (2006). Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health and Quality of Life Outcomes*, 4(1), 79.
- Field, A. P. (2005). *Discovering statistics using spss:(and sex, drugs and rock'n'roll)*. sage.
- Finch, W. H. (2011). A comparison of factor rotation methods for dichotomous data. *Journal of Modern Applied Statistical Methods*, 10(2), 14.
- Finney, S., & DiStefano, C. (2013). Dealing with nonnormality and categorical data in structural equation modeling. *A second Course in Structural Equation Modeling*. Greenwich, CT: Information Age.
- Fisher, R. A. (1954). *Statistical methods for research workers 12 th ed*. Oliver & Body.
- 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(4), 466.
- Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor analysis

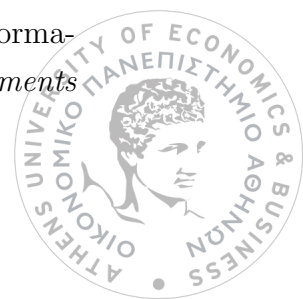


- with ordinal indicators: A monte carlo study comparing dwls and uls estimation. *Structural Equation Modeling*, *16*(4), 625–641.
- Foster, K. T., & Beltz, A. M. (2021). Heterogeneity in affective complexity among men and women. *Emotion*.
- Frison, L., & Pocock, S. J. (1992). Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. *Statistics in Medicine*, *11*(13), 1685–1704.
- Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*(4), 454.
- Gater, A., Nelsen, L., Coon, C. D., Eremenco, S., O'Quinn, S., Khan, A. H., ... others (2022). Asthma daytime symptom diary (adsd) and asthma nighttime symptom diary (ansd): Measurement properties of novel patient-reported symptom measures. *The Journal of Allergy and Clinical Immunology: In Practice*, *10*(5), 1249–1259.
- Gortler, R., Fox, J.-P., & Twisk, J. W. (2015). Why item response theory should be used for longitudinal questionnaire data analysis in medical research. *BMC Medical Research Methodology*, *15*(1), 1–12.
- Greenier, K. D., Kernis, M. H., McNamara, C. W., Waschull, S. B., Berry, A. J., Herlocker, C. E., & Abend, T. A. (1999). Individual differences in reactivity to daily events: Examining the roles of stability and level of self-esteem. *Journal of Personality*, *67*(1), 187–208.
- Greenland, S., & Finkle, W. D. (1995). A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American Journal of Epidemiology*, *142*(12), 1255–1264.
- Gregorich, S. E. (2006). Do self-report instruments allow meaningful comparisons across diverse population groups? testing measurement invariance using the confirmatory factor analysis framework. *Medical Care*, *44*(11 Suppl 3), S78.
- Griffiths, P., Williams, A., & Brohan, E. (2022). How do the number of missing daily diary days impact the psychometric properties and meaningful change thresholds arising from a weekly average summary score? *Quality of Life Research*, *31*(12), 3433–3445.
- Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15.
- Hamdan, M., & Martinson, E. (1971). Maximum likelihood estimation in the bivariate binomial (0, 1) distribution: application to  $2 \times 2$  tables. *Australian Journal of Statistics*, *13*(3), 154–158.
- Hardin, J. W., Hardin, J. W., Hilbe, J. M., & Hilbe, J. (2007). *Generalized linear*

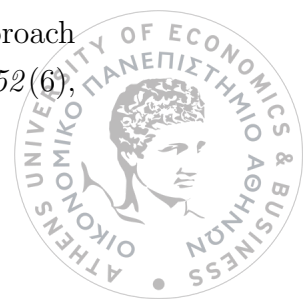


*models and extensions*. Stata press.

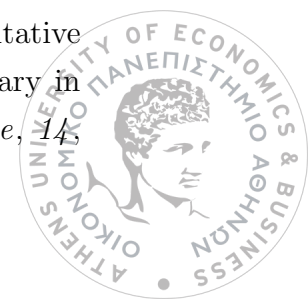
- Harman, H. H., & Fukuda, Y. (1966). Resolution of the heywood case in the minres solution. *Psychometrika*, *31*(4), 563–571.
- Härnqvist, K. (1978). Primary mental abilities at collective and individual levels. *Journal of Educational Psychology*, *70*(5), 706.
- Haven, S., & ten Berge, J. M. (1977). *Tucker's coefficient of congruence as a measure of factorial invariance: An empirical study*. Psychologische Instituten der Rijksuniversiteit Groningen.
- Heck, R. H. (2001). Multilevel modeling with sem. In *New developments and techniques in structural equation modeling* (pp. 109–148). Psychology Press.
- Higgins, J. P., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.
- Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, *42*(4), 609–629.
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367.
- Hooker, K. (1991). Change and stability in self during the transition to retirement: An intraindividual study using p-technique factor analysis. *International Journal of Behavioral Development*, *14*(2), 209–233.
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Evaluating model fit: a synthesis of the structural equation modelling literature. In *7th european conference on research methodology for business and management studies* (pp. 195–200).
- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179–185.
- Horstmann, K. T. (2021). Experience sampling and daily diary studies: Basic concepts, designs, and challenges. In *The handbook of personality dynamics and processes* (pp. 791–814). Elsevier.
- Houts, C. R., Morlock, R., Blum, S. I., Edwards, M. C., & Wirth, R. (2018). Scale development with small samples: A new application of longitudinal item response theory. *Quality of Life Research*, *27*(7), 1721–1734.
- Hox, J. (1998). Multilevel modeling: When and why. In *Classification, data analysis, and data highways* (pp. 147–154). Springer.
- Hox, J., & Maas, C. (2004). Multilevel structural equation models: The limited information approach and the multivariate multilevel approach. In *Recent developments on structural equation models* (pp. 135–149). Springer.



- Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.
- Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.
- Huang, F. L. (2017). Conducting multilevel confirmatory factor analysis using r. *Unpublished Manuscript*). <http://faculty.missouri.edu/huangf/data/mcfa/MCFA%20in%20R%20HUANG.pdf>.
- Huber, P. J. (1967). Under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability: Weather modification; university of california press: Berkeley, ca, usa* (p. 221).
- Jones, C. J., & Nesselroade, J. R. (1990). Multivariate, replicated, single-subject, repeated measures designs and p-technique factor analysis: A review of intraindividual change studies. *Experimental Aging Research, 16*(4), 171–183.
- Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8*(3), 325–352.
- Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141–151.
- Kaplan, D., & Elliott, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics, 22*(3), 323–347.
- Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*(3), 333–351.
- Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.
- Koch, G. G. (2004). Intraclass correlation coefficient. *Encyclopedia of statistical sciences*.
- Kolenikov, S., & Bollen, K. A. (2012). Testing negative error variances: Is a heywood case a symptom of misspecification? *Sociological Methods & Research, 41*(1), 124–167.
- Kurz, A. S., Johnson, Y. L., Kellum, K. K., & Wilson, K. G. (2019). How can process-based researchers bridge the gap between individuals and groups? discover the dynamic p-technique. *Journal of Contextual Behavioral Science, 13*, 60–65.
- Larsen, R. J. (1987). The stability of mood variability: A spectral analytic approach to daily mood assessments. *Journal of Personality and Social Psychology, 52*(6), 1195.

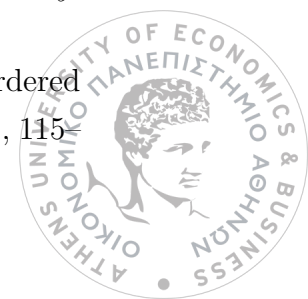


- Lebo, M. A., & Nesselroade, J. R. (1978). Intraindividual differences dimensions of mood change during pregnancy identified in five p-technique factor analyses. *Journal of Research in Personality*, *12*(2), 205–224.
- Lee, I. A., & Little, T. D. (2012). P-technique factor analysis. *American Psychological Association*.
- Lee, S.-Y., & Poon, W.-Y. (1985). Further developments on constrained estimation in analysis of covariance structures. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *34*(3), 305–316.
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.
- Lindqvist, K., Falkenström, F., Sandell, R., Holmqvist, R., Ekeblad, A., & Thorén, A. (2017). Multilevel exploratory factor analysis of the feeling word checklist–24. *Assessment*, *24*(7), 907–918.
- Lipton, R. B., Gandhi, P., Stokes, J., Cala, M. L., Evans, C. J., Knoble, N., . . . Dodick, D. W. (2022). Development and validation of a novel patient-reported outcome measure in people with episodic migraine and chronic migraine: The activity impairment in migraine diary. *Headache: The Journal of Head and Face Pain*, *62*(1), 89–105.
- Lischetzke, T., & Könen, T. (2020). Daily diary methodology. In *Encyclopedia of quality of life and well-being research* (pp. 1–8). Springer.
- Little, J. (2013). Multilevel confirmatory ordinal factor analysis of the life skills profile–16. *Psychological Assessment*, *25*(3), 810.
- Little, R. J. (1992). Regression with missing x's: a review. *Journal of the American Statistical Association*, *87*(420), 1227–1237.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130.
- MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84.
- Marchenko, V. A., & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, *114*(4), 507–536.
- Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing hu and bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341.
- Martin Nguyen, A., Bacci, E., Dicipinigaitis, P., & Vernon, M. (2020). Quantitative measurement properties and score interpretation of the cough severity diary in patients with chronic cough. *Therapeutic Advances in Respiratory Disease*, *14*



1753466620915155.

- Matthews, J., Altman, D. G., Campbell, M., & Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, *300*(6719), 230–235.
- McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 215–232.
- McNeish, D., Mackinnon, D. P., Marsch, L. A., & Poldrack, R. A. (2021). Measurement in intensive longitudinal data. *Structural Equation Modeling*, *28*(5), 807–822.
- Mehta, P. D., & Neale, M. C. (2005). People are variables too: multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259.
- Miettinen, O. S. (2012). Theoretical epidemiology: principles of occurrence research in medicine. In *Theoretical epidemiology: principles of occurrence research in medicine* (pp. 359–359).
- Mindrila, D. (2010). Maximum likelihood (ml) and diagonally weighted least squares (dwls) estimation procedures: A comparison of estimation bias with ordinal and multivariate non-normal data. *International Journal of Digital Society*, *1*(1), 60–66.
- Molenaar, P. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, *50*(2), 181–202.
- Molenaar, P. C. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218.
- Molenaar, P. C., & Nesselroade, J. R. (2009). The recoverability of p-technique factor analysis. *Multivariate Behavioral Research*, *44*(1), 130–141.
- Moskowitz, D. S., & Young, S. N. (2006). Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of Psychiatry and Neuroscience*, *31*(1), 13–20.
- Mueller, R. O. (1999). *Basic principles of structural equation modeling: An introduction to lisrel and eqs*. Springer Science & Business Media.
- Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430.
- Muraki, E., & Carlson, J. E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, *19*(1), 73–90.
- Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*(4), 551–560.
- Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115



132.

- Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal likert variables. *British Journal of Mathematical and Statistical Psychology*, *38*(2), 171–189.
- Muthén, B., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In *Multilevel analysis of educational data* (pp. 87–99). Elsevier.
- Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585.
- Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*(4), 338–354.
- Muthén, B. O. (1993). Goodness of fit with categorical and other nonnormal variables. *SAGE Focus Editions*, *154*, 205–205.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398.
- Muthén, B. O. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*.
- Napa Scollon, C., Prieto, C.-K., & Diener, E. (2009). Experience sampling: promises and pitfalls, strength and weaknesses. In *Assessing well-being* (pp. 157–180). Springer.
- Nesselroade, J. R., & Ram, N. (2004). Studying intraindividual variability: What we have learned that will help us understand lives in context. *Research in Human Development*, *1*(1-2), 9–29.
- Neyman, J. (1992). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. In *Breakthroughs in statistics* (pp. 123–150). Springer.
- Nguyen, T. H., Han, H.-R., Kim, M. T., & Chan, K. S. (2014). An introduction to item response theory for patient-reported outcome measurement. *The Patient-Patient-Centered Outcomes Research*, *7*(1), 23–35.
- Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.
- Omar, R. Z., Wright, E. M., Turner, R. M., & Thompson, S. G. (1999). Analysing repeated measurements data: a practical comparison of methods. *Statistics in Medicine*, *18*(13), 1587–1603.
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90.
- Ram, N., & Gerstorf, D. (2009). Time-structured and net intraindividual variability:



tools for examining the development of dynamic characteristics and processes. *Psychology and Aging*, 24(4), 778.

- R Core Team, R., et al. (2013). R: A language and environment for statistical computing.
- Reckase, M. D. (2009). Historical background for multidimensional item response theory (mirt). In *Multidimensional item response theory* (pp. 57–77). Springer.
- Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, 84(2), 126–136.
- Robinson, W. S. (2009). Ecological correlations and the behavior of individuals. *International Journal of Epidemiology*, 38(2), 337–341.
- Roesch, S. C., Aldridge, A. A., Stocking, S. N., Villodas, F., Leung, Q., Bartley, C. E., & Black, L. J. (2010). Multilevel factor analysis and structural equation modeling of daily diary coping data: Modeling trait and state variation. *Multivariate Behavioral Research*, 45(5), 767–789.
- Rosseel, Y. (2012). lavaan: An r package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.
- Sarstedt, M., & Wilczynski, P. (2009). More for less? a comparison of single-item and multi-item measures. *Die Betriebswirtschaft*, 69(2), 211.
- Savalei, V. (2014). Understanding robust corrections in structural equation modeling. *Structural Equation Modeling*, 21(1), 149–160.
- Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling*, 26(4), 576–590.
- Schermelleh-Engel, K., Moosbrugger, H., Müller, H., et al. (2003). Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online*, 8(2), 23–74.
- Schmidt, W., & Wisenbaker, J. (1986). *Hierarchical data analysis: An approach based on structural equations* (Tech. Rep.). Tech. Rep.
- Schneider, S., Junghaenel, D. U., Keefe, F. J., Schwartz, J. E., Stone, A. A., & Broderick, J. E. (2012). Individual differences in the day-to-day variability of pain, fatigue, and well-being in patients with rheumatic disease: associations with psychological variables. *Pain®*, 153(4), 813–822.
- Schneider, S., & Stone, A. A. (2016). Ambulatory and diary methods can facilitate the measurement of patient-reported outcomes. *Quality of Life Research*, 25(3), 497–506.
- Schoemann, A. M., Rhemtulla, M., & Little, T. D. (2014). Multilevel and longitudinal modeling.
- Schumacker, R. E., & Beyerlein, S. T. (2000). Confirmatory factor analysis with different correlation types and estimation methods. *Structural Equation Modeling*,



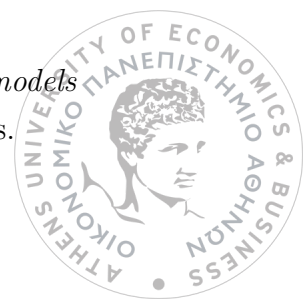
7(4), 629–636.

- Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. psychology press.
- Shi, D., & Maydeu-Olivares, A. (2020). The effect of estimation methods on sem fit indices. *Educational and Psychological Measurement, 80*(3), 421–445.
- Shi, D., Maydeu-Olivares, A., & Rosseel, Y. (2020). Assessing fit in ordinal factor analysis models: Srmr vs. rmsea. *Structural Equation Modeling, 27*(1), 1–15.
- Shoda, Y., Mischel, W., & Wright, J. C. (1994). Intraindividual stability in the organization and patterning of behavior: incorporating psychological situations into the idiographic analysis of personality. *Journal of Personality and Social psychology, 67*(4), 674.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multi-level, longitudinal, and structural equation models*. Chapman and Hall/CRC.
- Smyth, J. M., & Stone, A. A. (2003). Ecological momentary assessment research in behavioral medicine. *Journal of Happiness Studies, 4*(1), 35–52.
- Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. sage.
- Sokal, R. R. (1958). Thurstone's analytical method for simple structure and a mass modification thereof. *Psychometrika, 23*(3), 237–257.
- Song, H., & Zhang, Z. (2014). Analyzing multiple multivariate time series data using multilevel dynamic factor models. *Multivariate Behavioral Research, 49*(1), 67–77.
- Song, X.-Y., & Lee, S.-Y. (2003). Full maximum likelihood estimation of polychoric and polyserial correlations with missing data. *Multivariate Behavioral Research, 38*(1), 57–79.
- Spearman, C. (1961). " general intelligence" objectively determined and measured. *American Psychological Association*.
- Steiger, J. H. (1980). Statistically based tests for the number of common factors. In *Paper presented at the annual meeting of the psychometric society, iowa city, 1980*.
- Stevens, J. P. (2012). *Applied multivariate statistics for the social sciences*. Routledge.
- Stone, A. A., Broderick, J. E., & Kaell, A. T. (2010). Single momentary assessments are not reliable outcomes for clinical trials. *Contemporary Clinical Trials, 31*(5), 466–472.
- Stone, A. A., Broderick, J. E., Schneider, S., & Schwartz, J. E. (2012). Expanding options for developing outcome measures from momentary assessment data. *Psychosomatic Medicine, 74*(4), 387.
- Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics*



(Vol. 6). Pearson Boston, MA.

- Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.
- Tallis, G. (1962). The maximum likelihood estimation of correlation from contingency tables. *Biometrics*, *18*(3), 342–353.
- Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika*, *51*(4), 567–577.
- Thurstone, L. L. (1947). Multiple-factor analysis; a development and expansion of the vectors of mind.
- Trull, T. J., & Ebner-Priemer, U. (2014). The role of ambulatory assessment in psychological science. *Current Directions in Psychological Science*, *23*(6), 466–470.
- Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Tech. Rep.). Educational Testing Service Princeton Nj.
- Vach, W. (2012). *Logistic regression with missing values in the covariates* (Vol. 86). Springer Science & Business Media.
- van der Willik, E. M., Terwee, C. B., Bos, W. J. W., Hemmelder, M. H., Jager, K. J., Zoccali, C., . . . Meuleman, Y. (2021). Patient-reported outcome measures (proms): making sense of individual prom scores and changes in prom scores over time. *Nephrology*, *26*(5), 391–399.
- Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486–492.
- Van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika*, *43*, 225–243.
- Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. *Problems and Solutions in Human Assessment*, 41–71.
- Wessman, A. E., & Ricks, D. F. (1966). Mood and personality.
- Wirth, R., & Edwards, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, *12*(1), 58.
- Wu, J.-Y., & Kwok, O.-m. (2012). Using sem to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling*, *19*(1), 16–35.
- Xia, Y., & Yang, Y. (2019). Rmsea, cfi, and tli in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods*, *51*, 409–428.
- Yu, C.-Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. University of California, Los Angeles.



- Yuan, K.-H., & Bentler, P. M. (2007). 3. multilevel covariance structure analysis by fitting multiple single-level models. *Sociological Methodology*, *37*(1), 53–82.
- Zautra, A. J., Parrish, B. P., Van Puymbroeck, C. M., Tennen, H., Davis, M. C., Reich, J. W., & Irwin, M. (2007). Depression history, stress, and pain in rheumatoid arthritis patients. *Journal of Behavioral Medicine*, *30*, 187–197.
- Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. *Journal of Personality and Social Psychology*, *43*(1), 111.
- Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17*(2), 253–269.



# Appendix A

## Output tables



Table A.1: True loadings parameters, and EFA<sup>a</sup>estimated loadings within the second factor for the single selected day, weekly item average approaches, within- and between- individual analysis approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

		Scenario 1					Scenario 2				
		$L_T^c$	$L_M^d$			$L_T^c$	$L_d^b$				
$n^b$	Item		SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>		SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>
100	11	0.62	-0.12 - 1.00	0.43 - 0.84	-0.29 - 0.74	0.43 - 0.84	0.78	0.22 - 1.02	0.65 - 0.91	0.25 - 0.62	0.65 - 0.91
	12	0.69	0.16 - 0.92	0.72 - 0.98	-0.13 - 1.13	0.72 - 0.98	0.82	0.42 - 0.97	0.84 - 1.01	0.33 - 0.78	0.84 - 1.01
	13	0.66	0.20 - 1.00	0.63 - 0.91	-0.16 - 0.69	0.63 - 0.91	0.73	0.38 - 1.02	0.67 - 0.91	0.16 - 0.56	0.67 - 0.90
	14	0.66	0.05 - 0.98	0.67 - 0.95	-0.11 - 0.82	0.67 - 0.95	0.69	0.32 - 0.92	0.70 - 0.93	0.19 - 0.58	0.70 - 0.93
	15	0.70	0.19 - 0.95	0.60 - 0.93	-0.18 - 1.02	0.60 - 0.93	0.86	0.45 - 1.01	0.77 - 0.97	0.36 - 0.77	0.77 - 0.97
150	11	0.62	0.38 - 0.90	0.76 - 0.96	-0.08 - 0.60	0.76 - 0.96	0.78	0.54 - 0.99	0.83 - 0.99	0.32 - 0.67	0.83 - 0.99
	12	0.69	0.22 - 0.95	0.70 - 0.94	0.08 - 0.59	0.70 - 0.94	0.82	0.59 - 1.00	0.84 - 0.99	0.37 - 0.70	0.84 - 0.99
	13	0.66	0.34 - 0.97	0.68 - 0.93	-0.12 - 0.54	0.68 - 0.93	0.73	0.46 - 0.97	0.74 - 0.95	0.22 - 0.58	0.74 - 0.95
	14	0.66	0.21 - 0.87	0.68 - 0.94	-0.08 - 0.56	0.68 - 0.94	0.69	0.34 - 0.88	0.73 - 0.93	0.19 - 0.54	0.73 - 0.93
	15	0.70	0.35 - 0.94	0.75 - 0.96	-0.04 - 0.83	0.75 - 0.96	0.86	0.59 - 1.00	0.85 - 1.00	0.38 - 0.74	0.85 - 1.00
200	11	0.62	0.26 - 0.79	0.71 - 0.95	0.15 - 0.48	0.67 - 0.90	0.78	0.54 - 0.91	0.83 - 0.99	0.32 - 0.56	0.78 - 0.94
	12	0.69	0.31 - 0.88	0.76 - 0.96	0.23 - 0.62	0.75 - 0.94	0.82	0.57 - 0.93	0.84 - 0.99	0.40 - 0.66	0.83 - 0.98
	13	0.66	0.31 - 0.82	0.68 - 0.93	0.22 - 0.58	0.74 - 0.96	0.73	0.45 - 0.88	0.74 - 0.95	0.33 - 0.58	0.81 - 0.98
	14	0.66	0.31 - 0.84	0.68 - 0.94	0.23 - 0.59	0.74 - 0.94	0.69	0.35 - 0.84	0.73 - 0.93	0.27 - 0.56	0.78 - 0.94
	15	0.70	0.31 - 0.85	0.75 - 0.96	0.23 - 0.60	0.77 - 0.95	0.86	0.62 - 0.96	0.85 - 1.00	0.47 - 0.72	0.88 - 0.99
250	11	0.62	0.30 - 0.83	0.65 - 0.87	0.07 - 0.45	0.65 - 0.87	0.78	0.56 - 0.93	0.80 - 0.94	0.32 - 0.57	0.80 - 0.94
	12	0.69	0.34 - 0.89	0.75 - 0.93	0.19 - 0.61	0.75 - 0.93	0.82	0.54 - 0.93	0.82 - 0.94	0.34 - 0.60	0.82 - 0.94
	13	0.66	0.31 - 0.86	0.73 - 0.94	0.19 - 0.61	0.73 - 0.94	0.73	0.48 - 0.90	0.81 - 0.95	0.31 - 0.57	0.81 - 0.95
	14	0.66	0.25 - 0.85	0.66 - 0.87	0.14 - 0.55	0.66 - 0.87	0.69	0.35 - 0.84	0.67 - 0.86	0.17 - 0.46	0.67 - 0.86
	15	0.70	0.37 - 0.85	0.79 - 0.96	0.2 - 0.59	0.79 - 0.96	0.86	0.61 - 0.97	0.90 - 1.01	0.45 - 0.73	0.90 - 1.01
350	11	0.62	0.34 - 0.82	0.70 - 0.89	0.17 - 0.46	0.70 - 0.89	0.78	0.55 - 0.91	0.81 - 0.95	0.38 - 0.58	0.81 - 0.95



12	0.69	0.45 - 0.84	0.78 - 0.93	0.30 - 0.56	0.78 - 0.93	0.82	0.64 - 0.95	0.87 - 0.98	0.45 - 0.66	0.87 - 0.98
13	0.66	0.38 - 0.85	0.69 - 0.88	0.18 - 0.42	0.69 - 0.88	0.73	0.46 - 0.88	0.72 - 0.90	0.23 - 0.43	0.72 - 0.90
14	0.66	0.38 - 0.82	0.75 - 0.95	0.22 - 0.53	0.75 - 0.95	0.69	0.46 - 0.83	0.79 - 0.94	0.31 - 0.52	0.79 - 0.94
15	0.70	0.43 - 0.86	0.76 - 0.92	0.26 - 0.53	0.76 - 0.92	0.86	0.64 - 0.96	0.85 - 0.97	0.48 - 0.71	0.85 - 0.97

---

<sup>a</sup> Exploratory factor analysis

<sup>b</sup> Number of observations

<sup>c</sup> True value of loadings

<sup>d</sup> Model based estimated loadings

<sup>e</sup> Single selected day

<sup>f</sup> Weekly item average

<sup>g</sup> Within-individual analysis

<sup>h</sup> Between-individual analysis



Table A.2: True loadings parameters, and EFA<sup>a</sup>estimated loadings within the third factor for the single selected day, weekly item average approaches, within- and between-individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

n <sup>b</sup>	Item	Scenario 1					Scenario 2				
		$L_T^c$	$L_M^d$				$L_T^c$	$L_M^d$			
			SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>		SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>
100	16	0.63	0.18 - 0.98	0.66 - 0.98	-0.12 - 0.86	0.66 - 0.98	0.72	0.32 - 1.01	0.69 - 0.97	0.10 - 0.60	0.69 - 0.97
	17	0.62	0.18 - 0.98	0.66 - 0.98	-0.12 - 0.86	0.66 - 0.98	0.73	0.38 - 0.99	0.76 - 1.02	0.14 - 0.62	0.76 - 1.02
	18	0.69	0.18 - 1.06	0.70 - 0.99	-0.05 - 1.23	0.70 - 0.99	0.75	0.42 - 1.02	0.74 - 1.00	0.16 - 0.63	0.74 - 1.00
	19	0.70	0.17 - 0.97	0.72 - 1.01	-0.04 - 1.10	0.72 - 1.01	0.76	0.40 - 0.95	0.78 - 1.02	0.23 - 0.77	0.78 - 1.02
	20	0.65	0.16 - 0.98	0.66 - 0.96	-0.03 - 0.88	0.66 - 0.96	0.79	0.46 - 1.06	0.77 - 1.00	0.18 - 0.75	0.77 - 1.00
150	16	0.63	0.16 - 0.94	0.68 - 0.95	0.15 - 0.61	0.68 - 0.95	0.72	0.43 - 0.96	0.77 - 0.97	0.21 - 0.61	0.77 - 0.97
	17	0.62	0.22 - 0.95	0.70 - 0.94	0.08 - 0.59	0.70 - 0.94	0.73	0.59 - 1.00	0.84 - 0.99	0.37 - 0.70	0.84 - 0.99
	18	0.69	0.08 - 0.95	0.65 - 0.93	0.08 - 0.58	0.65 - 0.93	0.75	0.22 - 0.96	0.74 - 0.96	0.20 - 0.58	0.74 - 0.96
	19	0.70	0.22 - 0.98	0.75 - 0.99	0.10 - 0.84	0.75 - 0.99	0.76	0.39 - 0.96	0.78 - 1.00	0.26 - 0.70	0.78 - 1.00
	20	0.65	0.25 - 0.91	0.71 - 0.99	0.09 - 0.62	0.71 - 0.99	0.79	0.50 - 0.97	0.83 - 1.00	0.29 - 0.69	0.83 - 1.00
200	16	0.63	0.33 - 0.85	0.72 - 0.94	0.16 - 0.53	0.72 - 0.94	0.72	0.54 - 0.91	0.83 - 0.99	0.32 - 0.56	0.78 - 0.94
	17	0.62	0.31 - 0.88	0.76 - 0.96	0.23 - 0.62	0.75 - 0.94	0.73	0.50 - 0.92	0.79 - 0.98	0.28 - 0.58	0.79 - 0.98
	18	0.69	0.24 - 0.85	0.71 - 0.95	0.12 - 0.53	0.71 - 0.95	0.75	0.44 - 0.91	0.78 - 0.97	0.28 - 0.58	0.78 - 0.97
	19	0.70	0.43 - 0.91	0.77 - 0.97	0.19 - 0.66	0.77 - 0.97	0.76	0.55 - 0.91	0.82 - 0.97	0.32 - 0.64	0.82 - 0.97
	20	0.65	0.31 - 0.85	0.75 - 0.96	0.23 - 0.60	0.77 - 0.95	0.79	0.62 - 0.96	0.85 - 1.00	0.47 - 0.72	0.88 - 0.99
250	16	0.63	0.34 - 0.81	0.73 - 0.91	0.18 - 0.49	0.73 - 0.91	0.72	0.50 - 0.87	0.80 - 0.96	0.30 - 0.55	0.80 - 0.96
	17	0.69	0.39 - 0.81	0.70 - 0.94	0.22 - 0.53	0.70 - 0.94	0.73	0.52 - 0.90	0.79 - 0.97	0.34 - 0.61	0.79 - 0.97
	18	0.69	0.42 - 0.86	0.78 - 0.96	0.26 - 0.59	0.78 - 0.96	0.75	0.53 - 0.89	0.84 - 0.98	0.34 - 0.59	0.84 - 0.98
	19	0.70	0.43 - 0.86	0.77 - 0.95	0.22 - 0.56	0.77 - 0.95	0.76	0.53 - 0.89	0.81 - 0.97	0.31 - 0.57	0.81 - 0.97
	20	0.65	0.35 - 0.83	0.75 - 0.95	0.23 - 0.55	0.75 - 0.95	0.79	0.56 - 0.91	0.86 - 0.98	0.40 - 0.65	0.86 - 0.98
350	16	0.63	0.35 - 0.79	0.70 - 0.91	0.23 - 0.48	0.70 - 0.91	0.72	0.45 - 0.86	0.78 - 0.94	0.33 - 0.55	0.78 - 0.94



---

17	0.62	0.38 - 0.83	0.77 - 0.93	0.24 - 0.50	0.77 - 0.93	0.73	0.53 - 0.88	0.83 - 0.95	0.35 - 0.57	0.83 - 0.95
18	0.69	0.44 - 0.88	0.78 - 0.93	0.28 - 0.56	0.78 - 0.93	0.75	0.54 - 0.89	0.80 - 0.95	0.34 - 0.58	0.80 - 0.95
19	0.70	0.44 - 0.88	0.81 - 0.96	0.29 - 0.60	0.81 - 0.96	0.76	0.56 - 0.91	0.84 - 0.97	0.36 - 0.6	0.84 - 0.97
20	0.65	0.40 - 0.83	0.78 - 0.94	0.21 - 0.52	0.78 - 0.94	0.79	0.57 - 0.89	0.85 - 0.99	0.40 - 0.62	0.85 - 0.99

---

<sup>a</sup> Exploratory factor analysis

<sup>b</sup> Number of observations

<sup>c</sup> True value of loadings

<sup>d</sup> Model based estimated loadings

<sup>e</sup> Single selected day

<sup>f</sup> Weekly item average

<sup>g</sup> Within-individual analysis

<sup>h</sup> Between-individual analysis



Table A.3: True loadings parameters, and CFA<sup>a</sup>estimated loadings within the second factor for the single selected day, weekly item average approaches, within- and between- individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

n <sup>b</sup>	Item	Scenario 1					Scenario 2				
		$L_T^c$	$L_M^d$				$L_T^c$	$L_M^d$			
			SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>		SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>
100	11	0.62	0.18 - 1.00	0.51 - 0.82	0.14 - 0.46	0.71 - 1.38	0.78	0.21 - 0.96	0.70 - 0.88	0.34 - 0.61	0.80 - 1.11
	12	0.69	0.49 - 0.89	0.80 - 0.94	0.30 - 0.53	0.90 - 1.06	0.82	0.66 - 0.95	0.88 - 0.97	0.45 - 0.64	0.94 - 1.02
	13	0.66	0.43 - 0.93	0.75 - 0.89	0.22 - 0.49	0.87 - 1.07	0.73	0.52 - 0.94	0.79 - 0.91	0.28 - 0.51	0.89 - 1.04
	14	0.66	0.52 - 1.00	0.75 - 0.94	0.25 - 0.54	0.86 - 1.24	0.69	0.33 - 0.93	0.76 - 0.93	0.28 - 0.54	0.82 - 1.16
	15	0.70	0.49 - 0.96	0.67 - 0.91	0.24 - 0.62	0.79 - 1.17	0.86	0.46 - 1.08	0.83 - 0.96	0.45 - 0.72	0.89 - 1.08
150	11	0.62	0.35 - 0.86	0.75 - 0.92	0.25 - 0.52	0.88 - 1.10	0.78	0.60 - 0.94	0.88 - 0.96	0.42 - 0.63	0.94 - 1.04
	12	0.69	0.49 - 0.92	0.80 - 0.94	0.32 - 0.57	0.91 - 1.07	0.82	0.66 - 1.01	0.88 - 0.97	0.45 - 0.7	0.94 - 1.03
	13	0.66	0.43 - 0.93	0.75 - 0.89	0.22 - 0.49	0.87 - 1.07	0.73	0.52 - 0.94	0.79 - 0.91	0.28 - 0.51	0.89 - 1.04
	14	0.66	0.40 - 0.88	0.77 - 0.90	0.27 - 0.50	0.88 - 1.10	0.69	0.48 - 0.87	0.79 - 0.91	0.28 - 0.50	0.89 - 1.09
	15	0.70	0.46 - 0.92	0.80 - 0.93	0.29 - 0.56	0.92 - 1.09	0.86	0.69 - 1.01	0.90 - 0.98	0.49 - 0.67	0.95 - 1.04
200	11	0.62	0.32 - 0.82	0.75 - 0.92	0.25 - 0.48	0.86 - 1.07	0.78	0.59 - 0.90	0.88 - 0.96	0.41 - 0.59	0.91 - 1.01
	12	0.69	0.49 - 0.89	0.80 - 0.94	0.30 - 0.53	0.90 - 1.06	0.82	0.66 - 0.95	0.88 - 0.97	0.45 - 0.64	0.94 - 1.02
	13	0.66	0.38 - 0.83	0.75 - 0.89	0.27 - 0.51	0.92 - 1.07	0.73	0.55 - 0.86	0.79 - 0.91	0.37 - 0.57	0.93 - 1.05
	14	0.66	0.41 - 0.82	0.77 - 0.90	0.29 - 0.56	0.91 - 1.08	0.69	0.46 - 0.83	0.79 - 0.91	0.34 - 0.55	0.92 - 1.05
	15	0.70	0.50 - 0.87	0.80 - 0.93	0.33 - 0.54	0.90 - 1.06	0.86	0.71 - 0.97	0.90 - 0.98	0.53 - 0.69	0.95 - 1.02
250	11	0.62	0.35 - 0.83	0.68 - 0.84	0.19 - 0.41	0.80 - 1.05	0.78	0.63 - 0.92	0.83 - 0.91	0.37 - 0.53	0.90 - 1.01
	12	0.69	0.47 - 0.84	0.80 - 0.90	0.27 - 0.49	0.91 - 1.04	0.82	0.62 - 0.91	0.86 - 0.93	0.37 - 0.56	0.93 - 1.02
	13	0.66	0.46 - 0.80	0.81 - 0.91	0.27 - 0.48	0.92 - 1.06	0.73	0.55 - 0.85	0.85 - 0.93	0.36 - 0.54	0.94 - 1.03
	14	0.66	0.43 - 0.82	0.74 - 0.88	0.23 - 0.43	0.86 - 1.02	0.69	0.35 - 0.84	0.67 - 0.86	0.17 - 0.46	0.67 - 0.86
	15	0.70	0.48 - 0.85	0.83 - 0.92	0.31 - 0.51	0.91 - 1.04	0.86	0.74 - 0.98	0.92 - 0.97	0.54 - 0.69	0.96 - 1.02
350	11	0.62	0.42 - 0.81	0.71 - 0.84	0.24 - 0.42	0.86 - 1.06	0.78	0.63 - 0.90	0.84 - 0.92	0.40 - 0.55	0.92 - 1.03



---

12	0.69	0.53 - 0.87	0.83 - 0.92	0.34 - 0.54	0.93 - 1.03	0.82	0.70 - 0.93	0.91 - 0.96	0.50 - 0.64	0.96 - 1.02
13	0.66	0.42 - 0.86	0.73 - 0.86	0.24 - 0.42	0.84 - 1.01	0.73	0.55 - 0.87	0.78 - 0.88	0.29 - 0.45	0.87 - 1.00
14	0.66	0.49 - 0.79	0.80 - 0.91	0.32 - 0.49	0.92 - 1.04	0.69	0.52 - 0.80	0.82 - 0.90	0.35 - 0.50	0.92 - 1.03
15	0.70	0.54 - 0.86	0.77 - 0.89	0.31 - 0.50	0.89 - 1.03	0.86	0.73 - 0.96	0.88 - 0.94	0.50 - 0.66	0.93 - 1.01

---

<sup>a</sup> Confirmatory factor analysis

<sup>b</sup> Number of observations

<sup>c</sup> True value of loadings

<sup>d</sup> Model based estimated loadings

<sup>e</sup> Single selected day

<sup>f</sup> Weekly item average

<sup>g</sup> Within-individual analysis (the estimates are based on multilevel CFA)

<sup>h</sup> Between-individual analysis (the estimates are based on multilevel CFA)



Table A.4: True loadings parameters, and CFA<sup>a</sup>estimated loadings within the third factor for the single selected day, weekly item average approaches, within- and between-individual analysis approaches with the 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). a: Slope parameter from a multidimensional graded response model.

n <sup>b</sup>	Item	Scenario 1					Scenario 2				
		$L_T^c$	$L_M^d$				$L_T^c$	$L_M^d$			
			SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>		SD <sup>e</sup>	WIA <sup>f</sup>	WA <sup>g</sup>	BA <sup>h</sup>
100	16	0.63	0.35 - 0.95	0.74 - 0.91	0.15 - 0.50	0.87 - 1.12	0.72	0.52 - 1.02	0.79 - 0.92	0.22 - 0.54	0.88 - 1.04
	17	0.62	0.27 - 0.88	0.74 - 0.90	0.14 - 0.52	0.87 - 1.13	0.73	0.48 - 0.96	0.82 - 0.96	0.26 - 0.55	0.91 - 1.09
	18	0.69	0.41 - 0.94	0.79 - 0.95	0.24 - 0.60	0.89 - 1.08	0.75	0.54 - 0.96	0.84 - 0.96	0.30 - 0.60	0.91 - 1.05
	19	0.70	0.41 - 0.93	0.82 - 0.99	0.25 - 0.58	0.91 - 1.08	0.76	0.50 - 0.95	0.85 - 0.99	0.31 - 0.60	0.94 - 1.07
	20	0.65	0.26 - 0.97	0.76 - 0.93	0.19 - 0.58	0.85 - 1.11	0.79	0.49 - 1.00	0.84 - 0.95	0.32 - 0.64	0.89 - 1.05
150	16	0.63	0.44 - 0.94	0.75 - 0.91	0.21 - 0.56	0.87 - 1.09	0.72	0.56 - 0.93	0.82 - 0.94	0.30 - 0.57	0.91 - 1.05
	17	0.62	0.42 - 0.92	0.72 - 0.89	0.19 - 0.46	0.86 - 1.10	0.73	0.51 - 0.94	0.79 - 0.93	0.28 - 0.52	0.89 - 1.06
	18	0.69	0.44 - 0.87	0.81 - 0.94	0.28 - 0.54	0.91 - 1.06	0.75	0.56 - 0.92	0.86 - 0.95	0.34 - 0.59	0.93 - 1.05
	19	0.70	0.55 - 0.91	0.85 - 0.95	0.31 - 0.62	0.92 - 1.06	0.76	0.60 - 0.93	0.89 - 0.96	0.39 - 0.65	0.95 - 1.04
	20	0.65	0.46 - 0.89	0.81 - 0.94	0.27 - 0.53	0.90 - 1.07	0.79	0.63 - 0.95	0.89 - 0.97	0.42 - 0.68	0.95 - 1.03
200	16	0.63	0.43 - 0.83	0.79 - 0.92	0.26 - 0.49	0.90 - 1.09	0.72	0.54 - 0.88	0.86 - 0.94	0.35 - 0.55	0.93 - 1.04
	17	0.62	0.43 - 0.80	0.80 - 0.91	0.25 - 0.46	0.90 - 1.06	0.73	0.55 - 0.88	0.86 - 0.94	0.37 - 0.55	0.94 - 1.03
	18	0.69	0.53 - 0.89	0.82 - 0.93	0.31 - 0.55	0.92 - 1.05	0.75	0.61 - 0.91	0.85 - 0.94	0.37 - 0.59	0.93 - 1.04
	19	0.70	0.52 - 0.90	0.84 - 0.94	0.33 - 0.58	0.93 - 1.06	0.76	0.63 - 0.92	0.87 - 0.95	0.39 - 0.60	0.94 - 1.04
	20	0.65	0.50 - 0.86	0.81 - 0.92	0.28 - 0.51	0.90 - 1.06	0.79	0.67 - 0.94	0.88 - 0.95	0.41 - 0.60	0.94 - 1.03
250	16	0.63	0.43 - 0.84	0.78 - 0.90	0.27 - 0.48	0.90 - 1.05	0.72	0.57 - 0.89	0.85 - 0.93	0.35 - 0.53	0.93 - 1.03
	17	0.69	0.42 - 0.77	0.79 - 0.90	0.28 - 0.46	0.91 - 1.06	0.73	0.57 - 0.85	0.87 - 0.94	0.39 - 0.53	0.93 - 1.03
	18	0.69	0.50 - 0.85	0.84 - 0.94	0.34 - 0.54	0.92 - 1.05	0.75	0.62 - 0.88	0.88 - 0.95	0.39 - 0.56	0.96 - 1.03
	19	0.70	0.53 - 0.87	0.83 - 0.92	0.31 - 0.50	0.92 - 1.04	0.76	0.61 - 0.90	0.87 - 0.93	0.38 - 0.55	0.94 - 1.02
	20	0.65	0.48 - 0.82	0.81 - 0.92	0.28 - 0.50	0.92 - 1.05	0.79	0.67 - 0.91	0.90 - 0.95	0.46 - 0.61	0.95 - 1.03
350	16	0.63	0.44 - 0.78	0.78 - 0.88	0.28 - 0.45	0.90 - 1.04	0.72	0.57 - 0.86	0.84 - 0.91	0.37 - 0.50	0.93 - 1.02



17	0.62	0.46 - 0.78	0.80 - 0.89	0.28 - 0.47	0.91 - 1.05	0.73	0.59 - 0.88	0.87 - 0.93	0.40 - 0.55	0.94 - 1.03
18	0.69	0.53 - 0.83	0.82 - 0.91	0.34 - 0.50	0.92 - 1.02	0.75	0.60 - 0.87	0.86 - 0.92	0.40 - 0.54	0.94 - 1.02
19	0.70	0.54 - 0.84	0.85 - 0.93	0.37 - 0.55	0.94 - 1.03	0.76	0.63 - 0.87	0.89 - 0.94	0.42 - 0.57	0.95 - 1.02
20	0.65	0.44 - 0.82	0.81 - 0.90	0.27 - 0.47	0.93 - 1.03	0.79	0.66 - 0.90	0.89 - 0.94	0.43 - 0.58	0.95 - 1.02

---

<sup>a</sup> Confirmatory factor analysis

<sup>b</sup> Number of observations

<sup>c</sup> True value of loadings

<sup>d</sup> Model based estimated loadings

<sup>e</sup> Single selected day

<sup>f</sup> Weekly item average

<sup>g</sup> Within-individual analysis (the estimates are based on multilevel CFA)

<sup>h</sup> Between-individual analysis (the estimates are based on multilevel CFA)

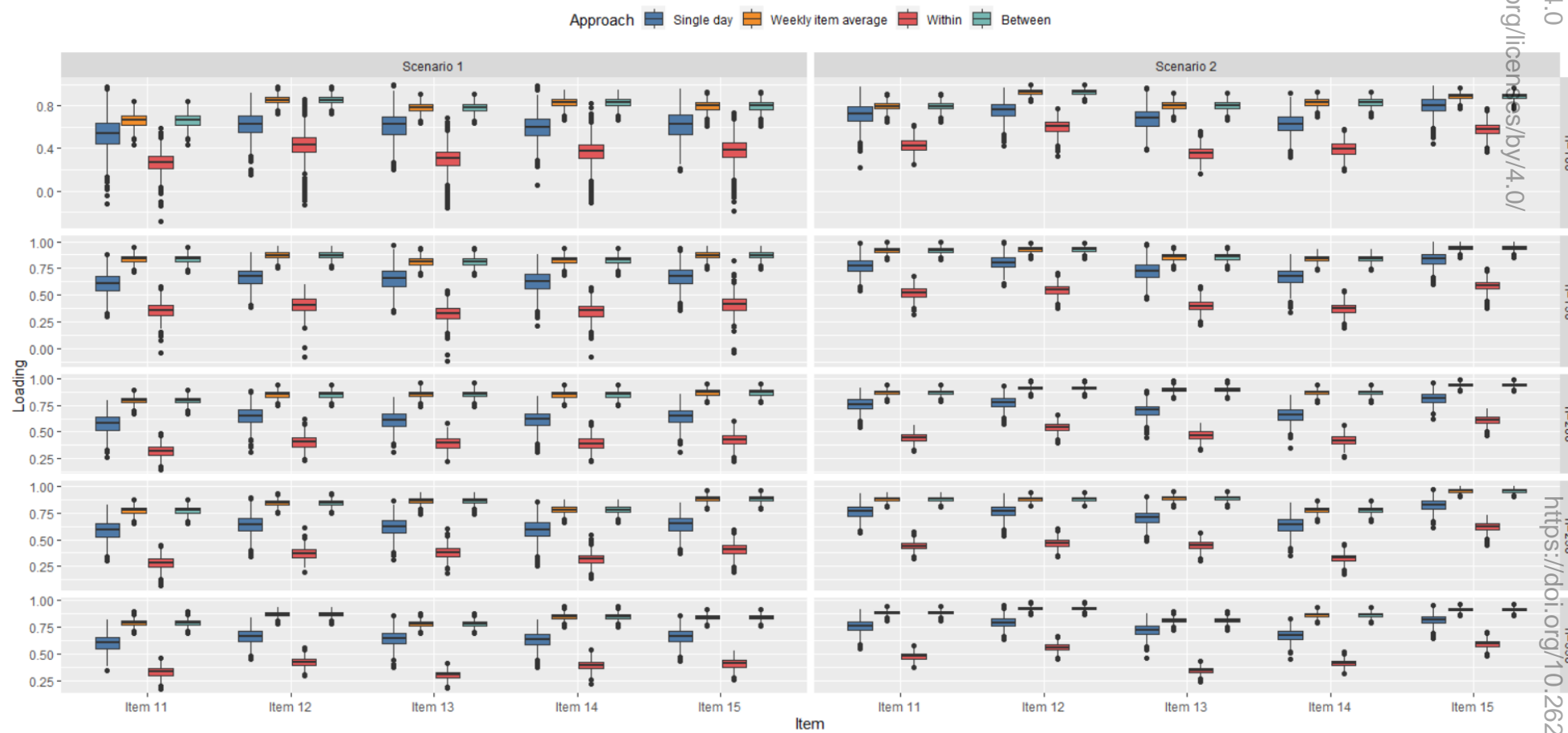


# Appendix B

## Output figures



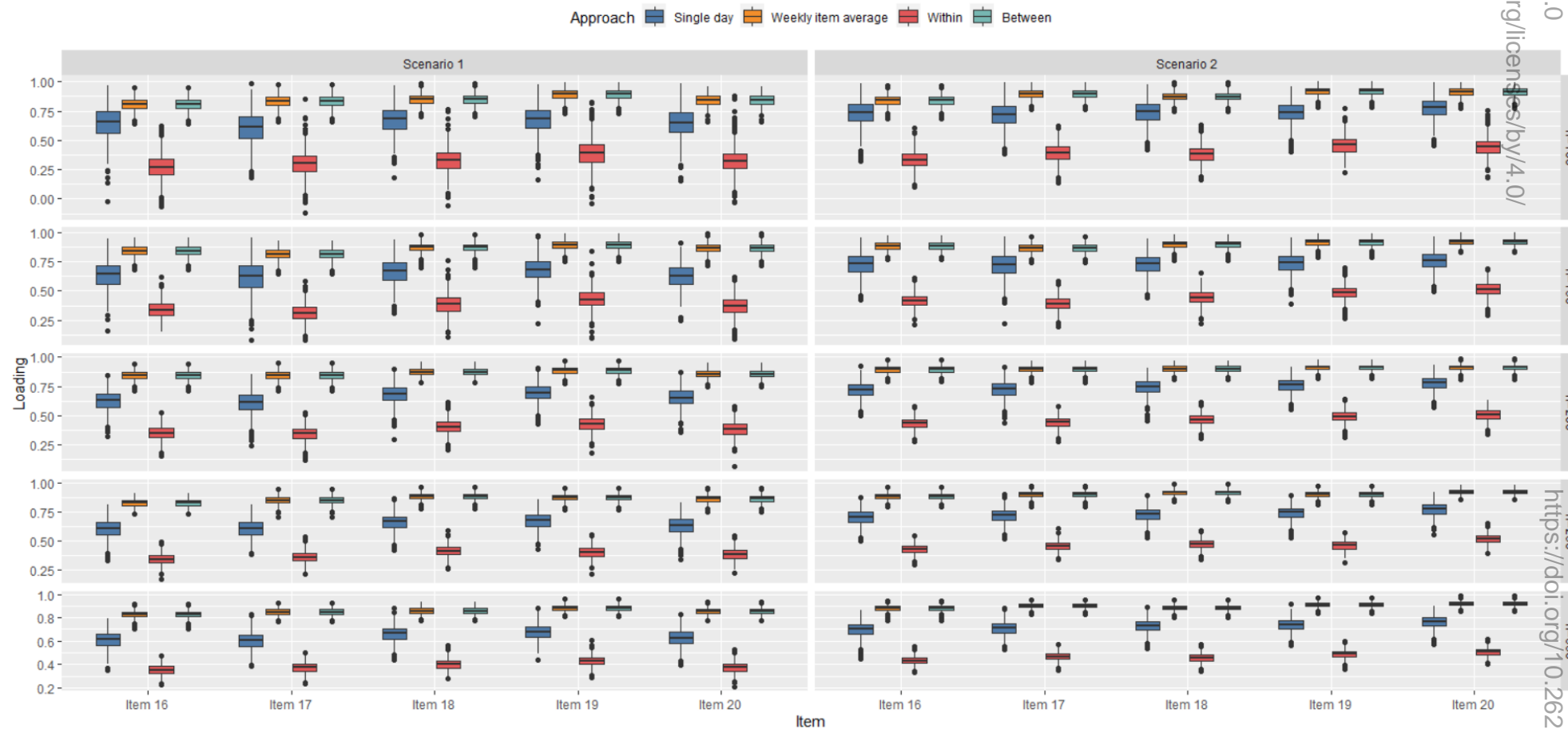
Figure B.1: Boxplot of the loadings within the second factor across 1,000 simulated datasets for EFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.



178



Figure B.2: Boxplot of the loadings within the third factor across 1,000 simulated datasets for EFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.



179



Figure B.3: Boxplot of the range of inter item observed and EFA-estimated correlation within the second factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.

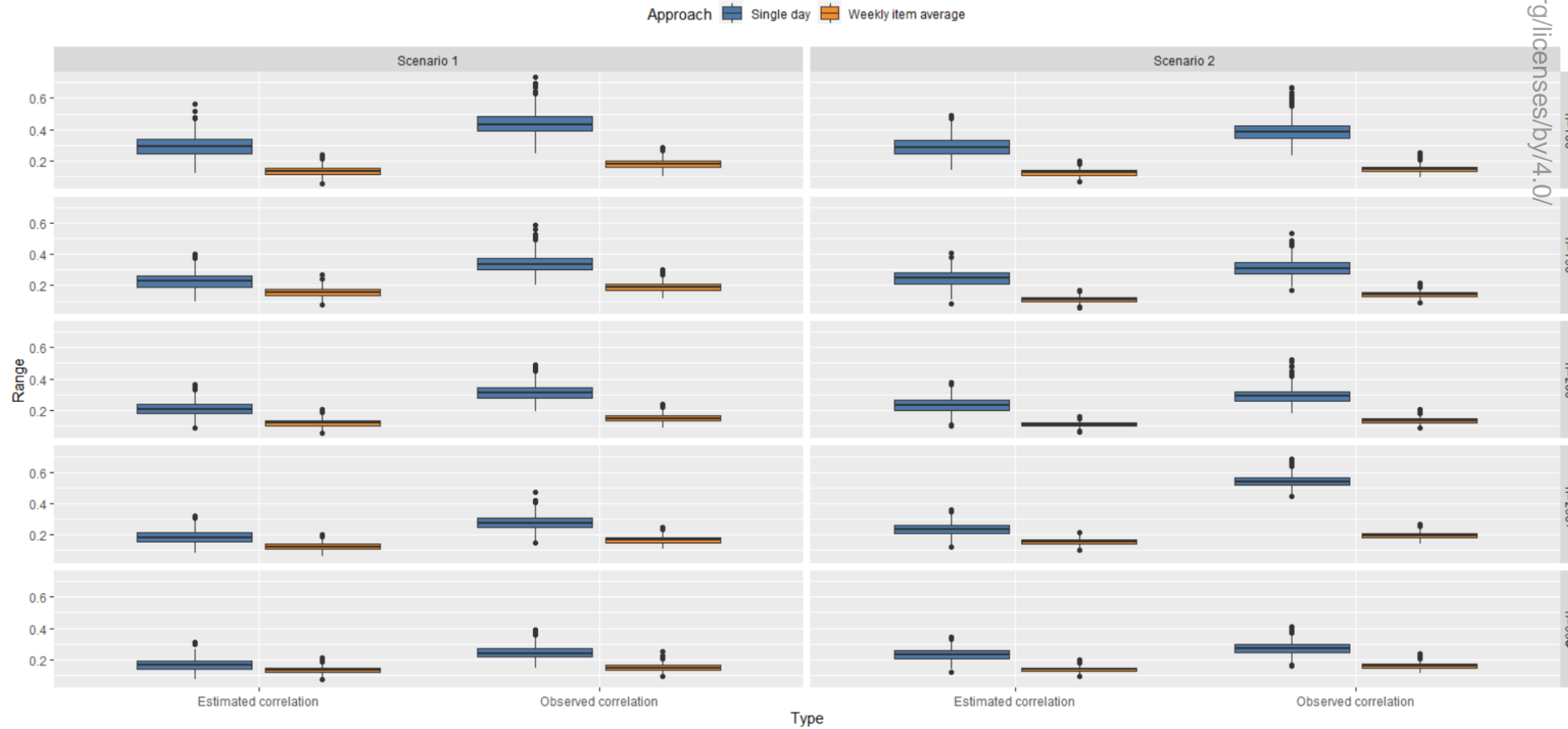
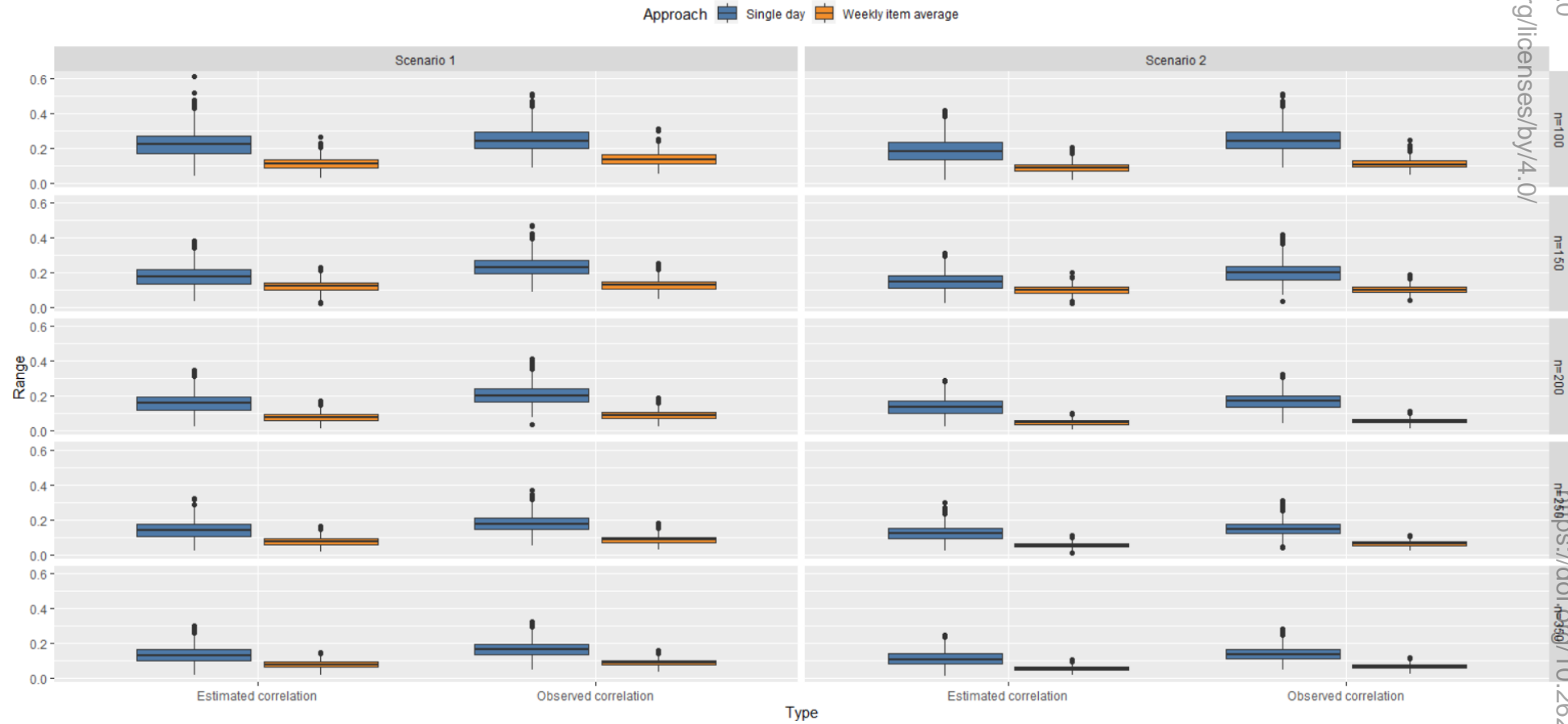


Figure B.4: Boxplot of the range of inter item observed and EFA-estimated correlation within the third factor across 1,000 simulated datasets for scenario 1 and 2 with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). EFA=Exploratory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.



n=100

n=150

n=200

n=250

n=350



Figure B.5: Boxplot of the loadings within the second factor across 1,000 simulated datasets for CFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.

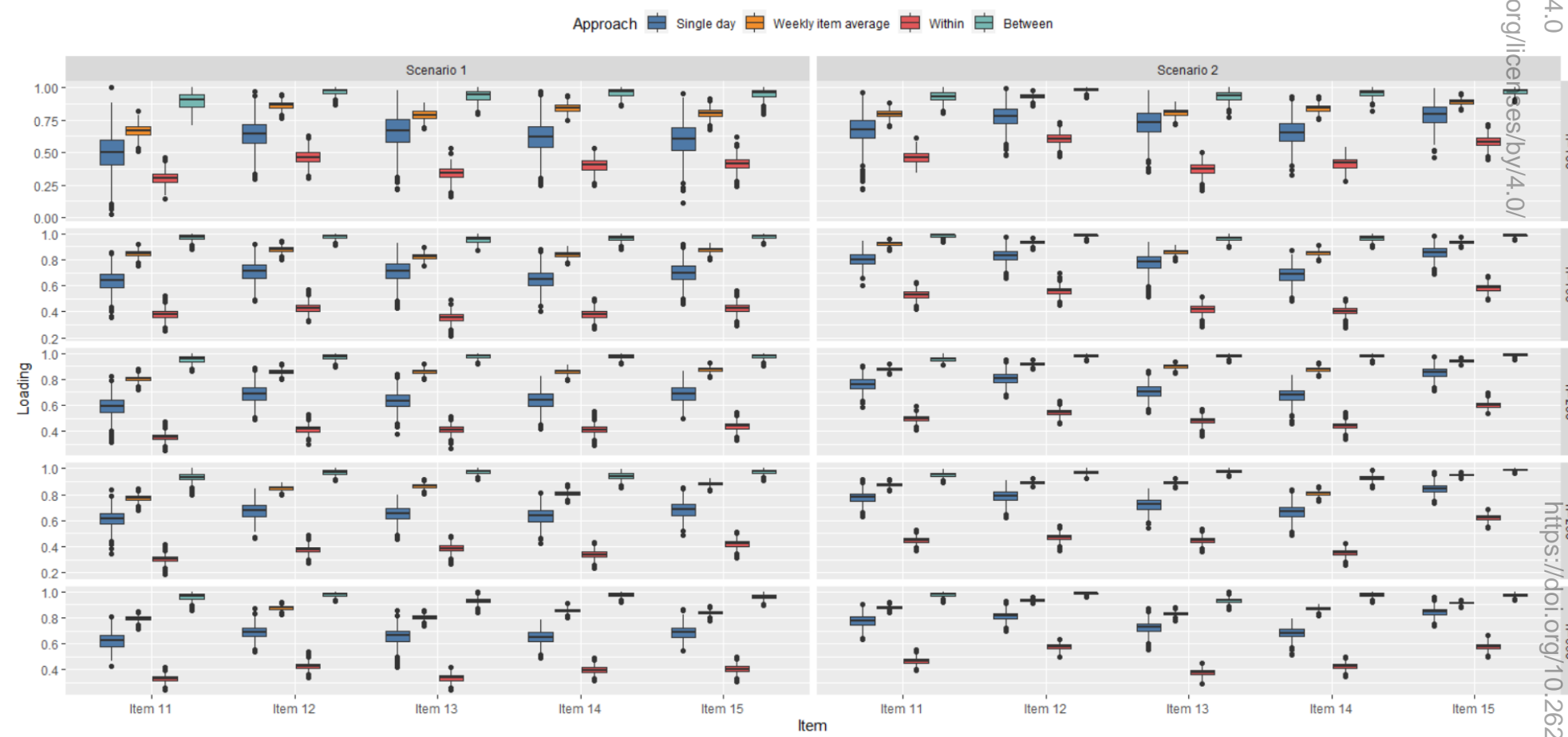


Figure B.6: Boxplot of the loadings within the third factor across 1,000 simulated datasets for CFA based on the different data handling approaches with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.

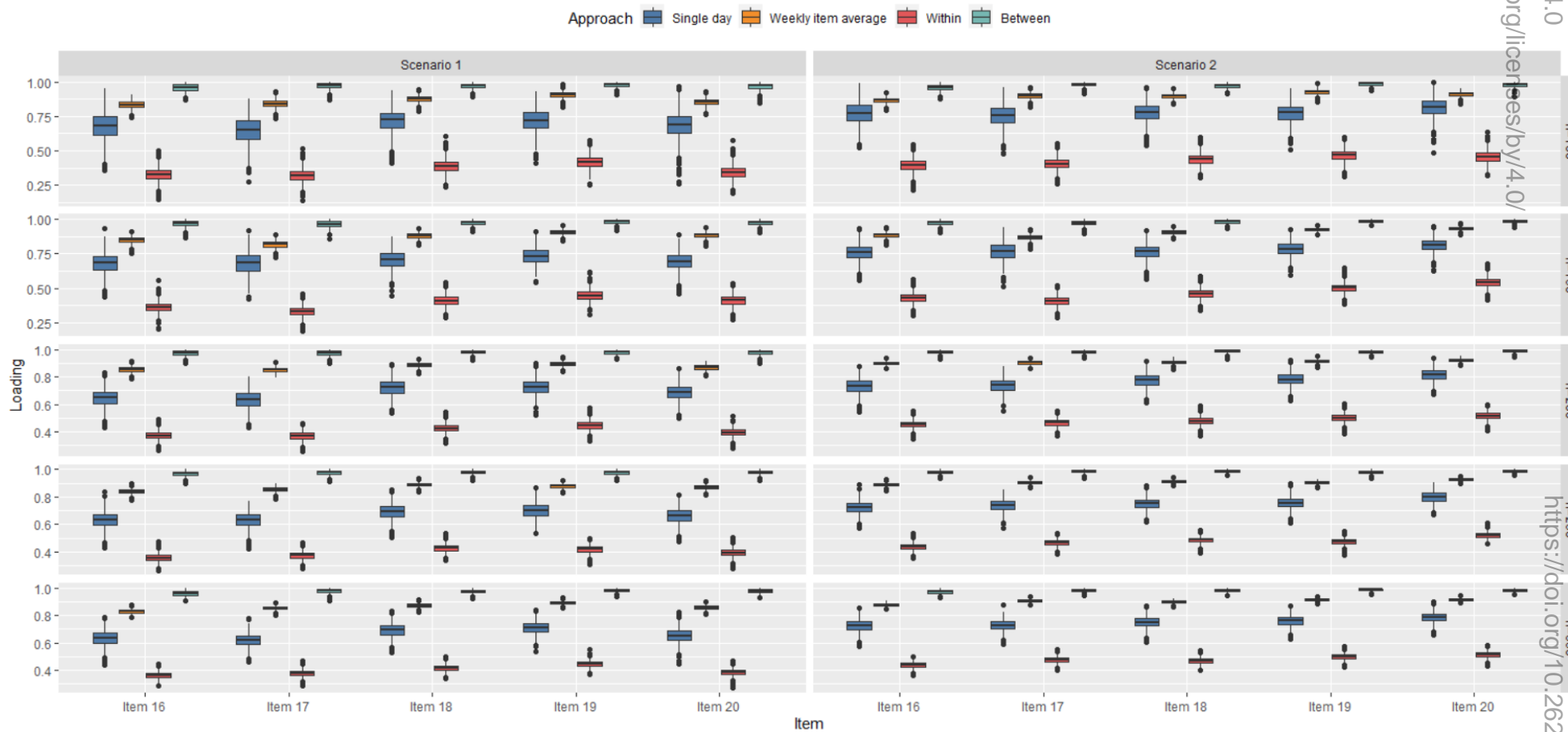


Figure B.7: Boxplot of the range of inter item observed and CFA-estimated correlation within the second factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis; a: Slope parameter from a multidimensional graded response model.

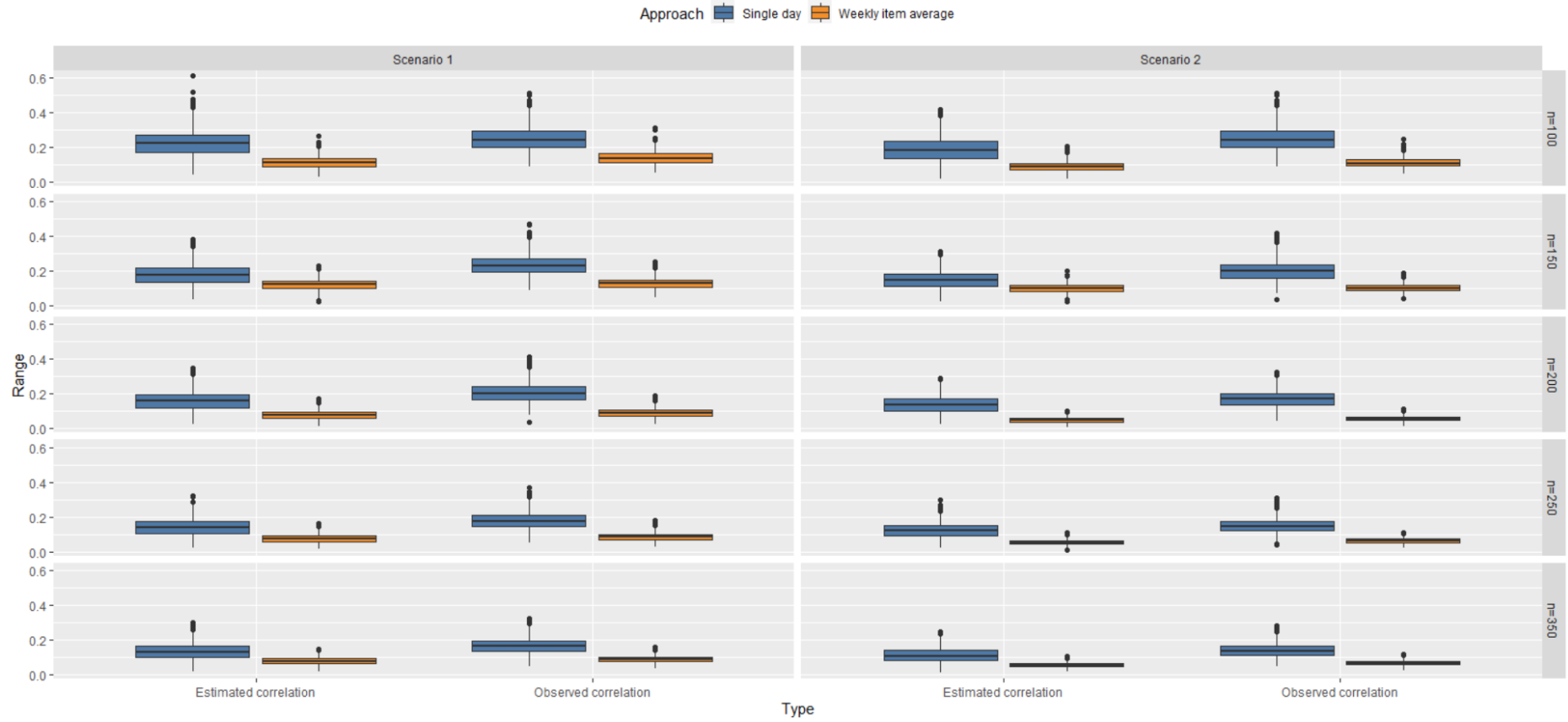
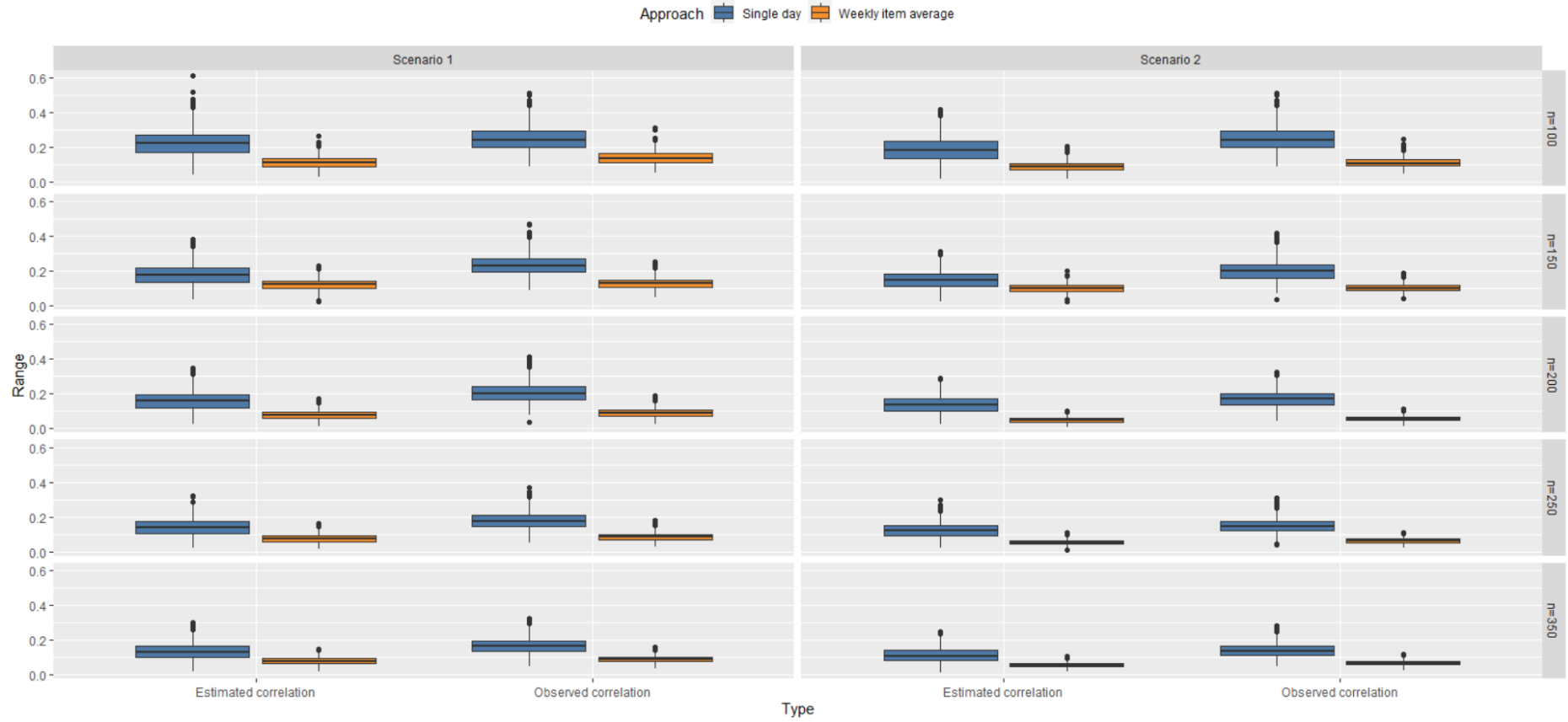


Figure B.8: Boxplot of the range of inter item observed and CFA-estimated correlation with the third factor across 1,000 simulated datasets with 100, 150, 200, 250 and 350 individuals across a 1-week period for scenario 1 ( $1.35 \leq a \leq 1.69$ ) and scenario 2 ( $a \geq 1.7$ ). CFA=Confirmatory factor analysis; Single day=Single selected day; Within=within-individual analysis; Between=between-individual analysis a: Slope parameter from a multidimensional graded response model.



# Appendix C

## Additional IRT models

For Rasch model, the difficulty parameter for each item and a single slope parameter are estimated. The model is characterized by its parsimony and this is the reason for which it could be also used even under the context of small sample size.

The model is described by the equation C.1:

$$P(Y_{ik} = 1|\theta_k) = \frac{\exp(1.7a(\theta_k - b_i))}{1 + \exp(1.7a(\theta_k - b_i))} \quad (\text{C.1})$$

where:

$\theta_k$  : Latent trait for k individual

$b_i$  : Difficulty parameter for item  $i$

1.7 : The scaling factor

$a$  : Slope parameter

$Y_{ik}$  : Response of individual k on item i

The 2 parameter logistic (2PL) model allows the estimation of difficulty and slope parameter for each item. The difference with the Rasch model is that it doesn't assume that the slope parameter is constant across all items. The model is described in C.2

$$P(Y_{ik} = 1|\theta_k) = \frac{\exp(a_i(\theta_k - b_i))}{1 + \exp(1 + a_i(\theta_k - b_i))} \quad (\text{C.2})$$



where:

$\theta_k$  : Latent trait for k individual

$b_i$  : Difficulty parameter for item  $i$

1.7 : The scaling factor

$a_i$  : Slope parameter for the  $i$  item

$Y_{ik}$  : Response of individual k on item  $i$

The 3-parameter logistic (3PL) model is used when a guessing parameter needs to be estimated. The model is described in equation C.3:

$$P(Y_{i,k} = 1|\theta_k) = c + (1 - c) \frac{\exp(a_i(\theta_k - b_i))}{1 + \exp(1 + a_i(\theta_k - b_i))} \quad (\text{C.3})$$

where:

$\theta_k$  : Latent trait for k individual

$b_i$  : Difficulty parameter for item  $i$

1.7 : The scaling factor

$a_i$  : Slope parameter for the  $i$  item

$Y_{ik}$  : Response of individual k on item  $i$

$c$  : Guessing parameter

