Fachbereich Erziehungswissenschaften und Psychologie der Freien Universität Berlin

Fachbereich Erziehungswissenschaften und Psychologie der Freien Universität Berlin

Latent Rater Agreement Models

Analyzing the Convergent and Discriminant Validity of Categorical Ratings

Dissertation
zur Erlangung des akademischen Grades
Doktor der Philosophie
(Dr. phil.)

vorgelegt von
Dipl. Psych.
Fridtjof Wilhelm Nußbeck

Erstgutachter: Prof. Dr. Michael Eid

Zweitgutachter: Prof. Dr. Michael Niedeggen

Datum der Disputation: 07. November 2008

BERLIN, 2008

Contents

1 INTRODUCTION – THE NEED FOR VALID MEASURES 5

2 MULTITRAIT-MULTIMETHOD MODELS AND RATER AGREEMENT MODELS 11

2.1 CONVERGENT AND DISCRIMINANT VALIDITY 11
2.2 THE LATENT MULTITRAIT-MULTIMETHOD MATRIX 15
2.3 MANIFEST RATER AGREEMENT MODELS 17
2.3.1 Rater Agreement Indices 18
2.3.2 Advantages and Limitations of Rater Agreement Indices 23
2.3.3 Rater Agreement Models 25
2.3.4 Advantages and Limitations of Manifest Rater Agreement Models 36

3 RESEARCH QUESTION 39

4 LATENT VARIABLE MODELS FOR CATEGORICAL DATA 42

4.1 LATENT VARIABLE MODELS FOR CATEGORICAL DATA 42
4.1.1 Formal Definition of the Log-Linear Model with Latent Variables 43
4.1.2 Estimation Process and Boundary Values 52
4.1.3 Application of the Log-Linear Model with One Latent Variable 54
4.2 EXTENSION TO MORE THAN ONE LATENT VARIABLE – CORRELATED TRAITS 67

5 LATENT RATER AGREEMENT MODELS 79

5.1 LATENT RATER AGREEMENT MODELS FOR STRUCTURALLY DIFFERENT RATERS 79
5.1.1 Definition of the Quasi-Independence I Model for Structurally Different Raters 82
5.1.2 Definition of the Quasi-Independence II Model for Structurally Different Raters 86
5.1.3 Definition of the Quasi-Symmetry Model for Structurally Different Raters 88
5.1.4 Applications of the Latent Rater Agreement Models for Structurally Different Raters 91
5.1.5 Comparison of the Latent Rater Agreement Models for Structurally Different Raters and Their Implications for the Analysis of Convergent and Discriminant Validity 105
5.2 LATENT RATER AGREEMENT MODELS FOR INTERCHANGEABLE RATERS 109
5.2.1 Measurement Invariance for Interchangeable Raters 110
5.2.2 Definition of the Quasi-Independence I Latent Rater Agreement Model for Interchangeable Raters 112
5.2.3 Definition of the Quasi-Independence II Latent Rater Agreement Model for Interchangeable Raters 113
5.2.4 Definition of the Symmetry (Saturated) Latent Rater Agreement Model for Interchangeable Raters 114
5.2.5 Applications of the Latent Rater Agreement Models for Interchangeable Raters 117
5.2.6 Implications of the Rater Agreement Models for Interchangeable Raters 122
5.3 DISCUSSION OF THE LATENT RATER AGREEMENT MODELS FOR STRUCTURALLY DIFFERENT AND INTERCHANGEABLE RATERS 124

6 CORRELATED TRAITS MULTITRAIT-MULTIRATER MODEL 126

6.1 DEFINITION OF THE CORRELATED TRAITS MULTITRAIT-MULTIRATER MODEL FOR STRUCTURALLY DIFFERENT RATERS 126
6.1.1 The Statistical Meaning of the Different Effects in the Saturated CT MTMR Model 128
6.2 THE CORRELATED TRAITS MULTITRAIT-MULTIRATER MODEL FOR INTERCHANGEABLE RATERS 150
6.2.1 Formal Representation of the Saturated CT MTMR Model for Interchangeable Raters 150
6.3 EMPIRICAL APPLICATIONS OF THE CT MTMR MODEL FOR STRUCTURALLY DIFFERENT AND INTERCHANGEABLE RATERS 169
6.3.1 Empirical Applications of the CT MTMR Model for Structurally Different Raters 170
6.3.2 Applications of the CT MTMR Models for Interchangeable Raters 192
6.4 DISCUSSION OF THE CT MTMR MODELS 204

7 SUMMARY AND DISCUSSION 206

7.1 SUMMARY OF THE MODEL PARAMETERS AND MODEL RESULTS 207
7.1.1 Latent Rater Agreement Models 207
7.1.2 Multitrait-Multirater (MTMR) Models 213

7.2 THE "JOINT FRAMEWORK" OF THE CT MTMR MODEL AND THE REALISTIC ACCURACY MODEL 219
7.3 LIMITATIONS OF THE MODELS AND FUTURE RESEARCH DIRECTIONS 224

REFERENCES 227

LIST OF FIGURES 237

LIST OF TABLES 238

APPENDIX A: DATA DESCRIPTION 243

A.1 FREQUENCY DISTRIBUTIONS OF THE BIG-FIVE ITEMS 244
A.2 RESPONSE PATTERNS OF THE SELF-REPORT DATA FOR NEUROTICISM 247

APPENDIX B: COLLAPSIBILITY THEOREM 249

APPENDIX C: LOG-LINEAR PARAMETERS OF THE LATENT RATER AGREEMENT MODELS FOR STRUCTURALLY DIFFERENT RATERS 250

C.1 SATURATED LATENT RATER AGREEMENT MODEL 250
C.2 INDEPENDENCE LATENT RATER AGREEMENT MODEL 252
C.3 QUASI-INDEPENDENCE I LATENT RATER AGREEMENT MODEL 256
C.4 QUASI-INDEPENDENE II LATENT RATER AGREEMENT MODEL 260
C.5 QUASI-SYMMETRY LATENT RATER AGREEMENT MODEL 264

APPENDIX E: LOGLINEAR PARAMETERS OF THE CT MTMR MODEL WITH TWO-VARIABLE EFFECTS AS HIGHEST ORDER INTERACTIONS 268

E.1: LOGLINEAR PARAMETERS OF THE MODEL FOR STRUCTURALLY DIFFERENT RATERS 268
E.2: LOGLINEAR PARAMETERS OF THE MODEL FOR INTERCHANGEABLE RATERS 270

APPENDIX F: INPUT FILES 272

F.1: INPUT FILES FOR RATER AGREEMENT MODELS 272
F.2: INPUT FILE FOR THE LOG-LINEAR MODELS WITH ONE LATENT VARIABLE AND THE MODEL WITH TWO LATENT VARIABLES 273
F.3: INPUT FILES FOR THE LATENT RATER AGREMENT MODELS 273
F.4: INPUT FILES FOR THE CT MTMR MODELS FOR STRUCTURALLY DIFFERENT AND INTERCHANGEABLE RATERS 277

APPENDIX G IN GERMAN LANGUAGE (ANHANG IN DEUTSCHER SPRACHE) 281

G.1: ZUSAMMENFASSUNG IN DEUTSCHER SPRACHE 281
G.2: LEBENSLAUF 285
G.3: ERKLÄRUNG 286

Acknowledgements

First of all, I would like to thank my advisor and supervisor Prof. Dr. Michael Eid. While I was a student at Trier University, he offered me a position as "studentische Hilfskraft" to work in the large MTMM study he was supervising at that time. The data I analyzed in my dissertation originated in this study. I followed him from Trier, to Landau, to Geneva, and to Freie University Berlin. He taught me to analyze MTMM data and strongly influenced my way of thinking about structural equation modeling. His friendly and cheerful character rendered the work atmosphere very positive and engaged me to continue on our common projects and on my dissertation.

I am also indebted to Christian Geiser and Dr. Delphine Courvoisier who gave me valuable hints and many critical remarks on large sections of my dissertation. Readers should also be thankful to them as well because they rendered many parts much more accessible to a general audience. Dr. Edith Braun contributed a great deal in streamlining the introduction and the second chapter of my dissertation. I would also like to thank Dr. Tanja Lischetzke who supervised me as "studentische Hilfskraft" at Trier University and who organized the data collection of the complicated MTMM study.

Many thanks also go to Prof. Dr. Michael Niedeggen who kindly agreed to be the second advisor for this dissertation. He also helped me to sustain the physically quite inactive time of writing a dissertation accompanying me during our long runs along the Teltow channel.

I also wish to thank our work group at Freie Universität Berlin. All members (Michael, Tanja L., Christian, Maike, Natalie, Martin, Claudia, Jana M., Jana H., Tanja K., Angela, Irina, and Luna) create a very sympathetic atmosphere, making it "more fun than work" to be at the University. During the final phase, Heike Bull supported me in providing loads of coffee during the late hours of the day keeping me at work when everybody else at Freie University was at home.

My mother, Prof. Dr. Susanne Nussbeck, did a great job not only during the phase of my dissertation but throughout my whole life. I am very grateful to her. She gave me the opportunity to conduct my studies at Trier University. She also did an exceptional job rendering many passages of this dissertation more accessible.

Finally, I would like to thank my partner Isabelle Staehli. She motivated me, helped me out of small crises, and assured me in very nervous phases just by being there. She did everything and even more than anyone could expect from her or his partner.

1 Introduction – The Need for Valid Measures

The validity and reliability of measures is of highest importance in many areas of psychology. Clinical judgments, for example, can have lasting consequences for clients. Invalid measurements bear risks like over– or underestimation of treatment effects, they may lead to the wrong diagnosis, they may indicate a suboptimal treatment, or, in the worst case, they might even not detect a relevant symptom at all. Burns and Haynes (2006) state that: "The validity of clinical research findings and clinical judgments depends on the validity of measures used in research and clinical activities" (p. 401). This is certainly true for all areas of psychology. In educational or developmental psychology, a newly developed schooling program may lead to disadvantages for children participating in this program against others who do not participate simply because an inadequate diagnostic tool is used. Traffic psychologists help to design road maps, crossings, and traffic lights to reduce the number of accidents. Therefore, they need valid diagnostic instruments to identify the best positions for them. All decisions in psychology should be based on the best information available. Information is best when it is objective, reliable, valid, and specific to a given problem (see e.g., Burns & Haynes, 2006; Courvoisier, Nussbeck, Eid, Geiser, & Cole, in press).

Psychological scales, measures, or ratings cannot be considered valid per se but their validity has to be proven in empirical applications. The general term "validation" (construct validation) subsumes many strategies that have been proposed to determine and improve the validity and reliability of psychological measures. Measures are unreliable when measurement error is large and invalid when systematic influences other than those one wants to measure have a strong impact on the measurement scores. We need to identify these influences to be sure that our measures truly measure what they are supposed to represent (e.g., Messick, 1995). If only those influences we wanted to capture are causes of the observed score[1] we may say that a score is valid:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment. […] Broadly speaking, then,

[1] The term score is used in its broadest sense. Any categorization and observation of consistent behaviors or attributes is conceived as a score.

> validity is an inductive summary of both the existing evidence and the potential consequences of score interpretation and use. (Messick, 1995, p. 13)

Validity and validation are, thus, at the core of scientific and applied psychology. There is an ongoing debate about the concept of validity. Some researchers say that it is a single property of scores and that these scores are valid or not in measuring an existing construct (see e.g., Borsboom, Mellenbergh, & Van Heerden, 2003, 2004). The link function relating the observed scores to the underlying construct is at the centre of this conceptualization of validity. Others explicitly refer to different types of validity that can be present to a certain degree (see e.g., Campbell & Fiske, 1959; Messick, 1995; Shadish, Cook, & Campbell, 2002). In this conceptualization, the nomological net is at the heart of validity. Scores are considered valid if they fit into a nomological net (show convergent and discriminant validity). I will refer to the latter concept of validity in this thesis. Three “types” of validity of a specific measure can be determined by one or all of the three main “types” of validation procedures (see Messick, 1995):

Content validity is examined by analyzing if the content of the test situation matches the area about which conclusions are to be drawn. Testing the knowledge of the Latin alphabet asking participants to type and name the different letters, for example, is highly content valid, because the area (the Latin alphabet) is well represented.

Criterion related validity is given when the score is highly associated with one or more external variables (criteria) that are considered to be related to the psychological construct. The criteria can be measured in the same situation (*concurrent validity*) or in future situations (*predictive validity*). An intelligence test may be highly criterion valid if it highly correlates with school achievement (for a conceptualization of intelligence close to academic skills).

Construct validity as in parts examined by the Multitrait–Multimethod Matrix (Campbell & Fiske, 1959) is concerned with the attributes (qualities) of a score. It is analyzed, which qualities are measured by a given score—that is, which concepts account for the performance on the test score. Some aspects of a given score can be determined by studying the association of the test score with other scores that are akin to the first score (*convergent validity*) and with scores that are supposed to measure completely different psychological constructs (*discriminant validity*). All items representing the same facet of an intelligence test as well as the results of different intelligence tests should be highly positively associated (convergent validity) because they are supposed to measure the same trait. For instance, there should be no or only a small association of scales measuring extraversion or neuroticism (high discriminant validity) because these traits are considered to be independent from each other.

The analysis of convergent and discriminant validity as done by the Multitrait–Multimethod (MTMM) matrix (Campbell & Fiske, 1959) has become one of the most important approaches for test-validation. Modern approaches of this analysis strategy offer the possibility to determine the reliability of multiple items representing one construct, the convergence of different methods measuring the same construct, the discriminant validity of different measures of different constructs measured by the same method, the influences due to method-specific effects, and to separate measurement error from true-scores.

So far, MTMM models have only been developed for the analysis of models with metric response variables or for variables with ordered response categories. To my knowledge, no MTMM model for response variables with non-ordered categories has been proposed so far. Almost all MTMM models that have been defined imply bivariate relationships between variables. That is, correlations or factor structures linking one manifest variable to its underlying latent variables. The latent variables in the structural part of the model are also associated via bivariate relationships.

In principle, these MTMM models assume linear relation-ships between latent variables. If two variables are positively correlated to each other, there must observational units that have small values on these two variables and other units that have high values on these two variables. The relationship can be considered “constant” (linear). Yet, relationships between variables do not have to be “constant” across all categories. Imagine the case with two distinct categorical variables consisting of three categories each. Principally none of the category combinations (elements of the joint distribution) is largely overrepresented compared to the expectancies given independence except for the joint categorization of $X = 2$ with $Y = 2$ (see Table 1.1.1). Therefore, the two variables are associated but the association originates in the overrepresentation of one particular cell combination (is not “constant” across all combinations). The latter piece of information is generally represented in models for categorical data because these models consist of parameters reflecting over- and underrepresentations of proportions (frequencies) of specific categories or category combinations. However, this piece of information is not directly available in the MTMM models proposed so far. It would be worthwhile to gather this information to examine, for example, if high convergent validity is due to an association between variables originating in systematic over- and underrepresentation of a large number of category combinations or in an over- and / or underrepresentation of only one or a few category combinations. Imagine physicians who rate radiographs. They may not show over- or underrepresented ratings with respect to the different pathologies but only with respect to normal (non-pathological) radiographs. Models analyzing the association of the two physicians’ ratings would indicate associated ratings although there is no overrepresentation (compared to independence) with respect to the pathological radiographs but only with respect to the normal cases. Models analyzing the category-specific over- or underrepresentation would allow for a more fine graded analysis indicating upon which cases the physicians agree and for which cases of pathologies they do not agree implying that they should improve their rating skills.

Table 1.1.1

Artificial frequency table of two categorical variables

			Y		
		1	2	3	
	1	15	15	15	45
X	2	15	60	15	90
	3	15	15	15	45
		45	90	45	180

Note. *X* and *Y* represent two distinct observed variables.

The aim of this dissertation is to define MTMM models for categorical outcomes. These models may help to understand more about the associations between different constructs because they principally allow for an examination which categories of different constructs are under- or overrepresented and for an integration of higher order interactions. These interactions depict the association of three or more constructs. The association of two constructs may change depending on the other construct. Highly extraverted individuals, for example, may more frequently be congruently judged as friendly and helpful by peer raters (high convergent validity) than highly introverted individuals upon whom the same raters do not agree or disagree more frequently than could be expected by chance (low convergent validity). Convergent and discriminant validity may therefore change as a function of the categories that are examined.

Models will be defined that allow for this kind of analyses. I will consider the special case of raters as methods (see Kenny, 1995); however, the results may be generalized to other methods in a straightforward way. In particular, the development of different Multitrait-Multirater (MTMR) models for non-ordered categorical data will be done in several steps. In Section 2.1, the concepts of convergent and discriminant validity will be defined and explained. The analysis of convergent and discriminant validity will be outlined with respect to the latent MTMM matrix (2.2).Since I will focus on raters as a special case of methods in the context of MTMM research (see e.g., Kenny, 1995) existing indices and models for the analysis of rater agreement will be revised in Section 2.3. This will lead to the research questions presented in Section 3.

In Section 4.1, the log-linear model with latent variables will be introduced. I will show how the model is defined and how to interpret the model parameters in a theoretically meaningful way. The model will be illustrated by an empirical application. In Section 4.2, the model will be extended to more than two latent variables providing the base for the definition of latent rater agreement models.

In Section 5, the latent rater agreement models will be defined on the latent level. These models allow analyzing the convergence of different raters with respect to different typologies. In Section 5.1, the latent rater agreement models for structurally different (heterogeneous) raters will be defined. The meaning of the model parameters will be explained in detail. Empirical applications serve to illustrate the models.

In Section 5.2, the previously defined latent rater agreement models will be defined for interchangeable (homogeneous) raters. The distinction of structurally different and interchangeable raters has severe consequences for the model definition with respect to the measurement models and the interaction terms. These differences will be outlined. Empirical applications illustrate the models.

In Section 6, the logic of rater agreement models will be combined with the strength of MTMM models allowing for the analysis of convergent and discriminant validity. I will explicitly refer to the criteria formulated by Campbell and Fiske (1959) to illustrate the strength of the newly developed models. The models allow for analyzing category-specific agreement rates (convergent validity), the discriminant validity between particular latent categories as well as a detailed analysis of (category-specific) rater specific effects. These effects reflect some of the determinants and moderators for rater accuracy models introduced by Funder (1995). In Section 6.1, these models will be defined for the case of structurally different raters. The case of interchangeable raters will be treated in Section 6.2. Empirical applications serve to illustrate the models.

Finally, the models will be discussed with respect to their theoretical implications for assessing the convergent and discriminant validity. Furthermore, it will be discussed how they can reflect complex effects of different latent categories across traits and across raters on each other, which may reveal important information about sources of agreement and disagreement. Future research directions as possible extensions to more than two or three traits, for example, will be discussed, too. Moreover, the newly developed models will be related to the rater accuracy model (Funder, 1995).

2 Multitrait-Multimethod Models and Rater Agreement Models

2.1 Convergent and Discriminant Validity

In their groundbreaking work “Convergent and discriminant validation by the multitrait-multimethod matrix” Campbell and Fiske (1959) proposed the multitrait-multimethod (MTMM) matrix as a methodological tool for test validation. More than 2000 citations during the first 33 years since published (Sternberg, 1992) and more than 4.500 citations until 2008[2] demonstrate the strong impact of Campbell and Fiske’s work. The initial analysis of the MTMM matrix with respect to the convergent and discriminant validity can be summarized in four points:

1. *Convergent validity* is given if different and independent measurement procedures or measures of the same construct converge. In general, measures are said to converge if they show sufficiently high correlations with each other. A valid score is a score which is reliably measured and whose systematic influences mainly correspond to the construct one wants to measure.
2. *Discriminant validity* is given if observed scores aiming at measuring distinct constructs do not converge. The scores of scales or other measurement procedures of one construct should show low correlations with scores measuring another construct.
3. *Trait-Method-Units* (TMU) are at the core of measurement. Each and every score in the behavioral sciences depends on influences due to the construct (*trait*) and properties of the measurement method (*method*). Method has become a term with a widespread meaning: A method may represent scales, raters, items, parcels of a test, measurement situations (e.g., field vs. laboratory), or occasions of measurement. Biesanz and West (2004), for example, give an overview of the meaning of the term “method” in modern psychometric models. Burns and Haynes (2006) identify different sources of variance of clinical measures that may all be modeled as methods in the sense of Campbell and Fiske. In this contribution, I will only consider raters as a specific method. Raters are one of the most common types of methods applied in psychology (see Kenny, 1995).

[2] information retrieved from isi web of knowledge (http://apps.isiknowledge.com) on March 26, 2008.

4. *More than one trait and more than one method (rater)* are needed to separate influences due to trait and method effects. More than one method is needed to identify the influence of the trait (construct). High correlations of different measures representing the same trait originating in different methods indicate the influence of the trait. More than one trait is needed to identify the influences due to the different methods. Correlations of measures belonging to the same method but different traits indicate method-specific influences.

Relying on these four considerations Campbell and Fiske (1959) introduced the Multitrait-Multimethod (MTMM) matrix (see Table 2.1.1). This matrix consists of the correlations between all trait scores measured with different methods. Additionally, the reliability can be depicted on the main diagonal. In this matrix, Campbell and Fiske identified four different key components for determining the convergent and discriminant validity. The four components can be found in two different blocks of the MTMM matrix.

Monomethod blocks are the cells combining scores of different traits measured by one single method (the method remains the same: $M_j = M_{j'}$). In these *monomethod* blocks, the *reliability* estimates (*monotrait-monomethod correlations*; depicted with R^2) and the *heterotrait-monomethod* correlations $\left(\text{e.g., } r_{(T_2M_1,T_1M_1)}\right)$ can be found (grey shaded triangles). Heterotrait-monomethod correlations represent the association of two distinct traits measured by one method. In general, these correlations should be rather low. However, these correlations represent influences due to the theoretically expected association of the two constructs but also influences due to the specific method. In the case of different raters as methods, these correlations are influenced by the association of the two traits, say openness and extraversion, and also by the rater-specific view of this association (e.g., the presence of a halo-effect may lead to an overestimation of the correlation of openness and extraversion).

In the *heteromethod* blocks, two types of correlations can be found. These correlations indicate the *convergent validity* and *heterotrait-heteromethod* correlations. Convergent validity (*monotrait-heteromethod correlations*) can be found on the *validity diagonals* between the triangles. These correlations $\left(r_{(T_1M_2,T_1M_1)}; r_{(T_2M_2,T_2M_1)}; r_{(T_3M_2,T_3M_1)}, \text{for example}\right)$ depict the convergence of trait measures measured by different methods (1 and 2 in the example).

Table 2.1.1

Multitrait-Multimethod matrix for three traits (T_1, T_2, and T_3) measured by three methods (M_1, M_2, and M_3)

			M_1			M_2			M_3	
		T_1	T_2	T_3	T_1	T_2	T_3	T_1	T_2	T_3
M_1	T_1	R^2								
	T_2	$r_{(T_2M_1,T_1M_1)}$	R^2							
	T_3	$r_{(T_3M_1,T_1M_1)}$	$r_{(T_3M_1,T_2M_1)}$	R^2						
M_2	T_1	$r_{(T_1M_2,T_1M_1)}$	$r_{(T_1M_2,T_2M_1)}$	$r_{(T_1M_2,T_3M_1)}$	R^2					
	T_2	$r_{(T_1M_2,T_2M_1)}$	$r_{(T_2M_2,T_2M_1)}$	$r_{(T_2M_2,T_3M_1)}$	$r_{(T_2M_2,T_1M_2)}$	R^2				
	T_3	$r_{(T_3M_2,T_1M_1)}$	$r_{(T_3M_2,T_2M_1)}$	$r_{(T_3M_2,T_3M_1)}$	$r_{(T_3M_2,T_1M_2)}$	$r_{(T_3M_2,T_2M_2)}$	R^2			
M_3	T_1	$r_{(T_1M_3,T_1M_1)}$	$r_{(T_1M_3,T_2M_1)}$	$r_{(T_1M_3,T_3M_1)}$	$r_{(T_1M_3,T_1M_2)}$	$r_{(T_1M_3,T_2M_2)}$	$r_{(T_1M_3,T_3M_2)}$	R^2		
	T_2	$r_{(T_2M_3,T_1M_1)}$	$r_{(T_2M_3,T_2M_1)}$	$r_{(T_2M_3,T_3M_1)}$	$r_{(T_2M_3,T_1M_2)}$	$r_{(T_2M_3,T_2M_2)}$	$r_{(T_2M_3,T_3M_2)}$	$r_{(T_2M_3,T_1M_3)}$	R^2	
	T_3	$r_{(T_3M_3,T_1M_1)}$	$r_{(T_3M_3,T_2M_1)}$	$r_{(T_3M_3,T_3M_1)}$	$r_{(T_3M_3,T_1M_2)}$	$r_{(T_3M_3,T_2M_2)}$	$r_{(T_3M_3,T_3M_2)}$	$r_{(T_3M_3,T_1M_3)}$	$r_{(T_3M_3,T_2M_3)}$	R^2

Note. R^2: Reliability; $r_{(T_iM_j,T_{i'}M_{j'})}$: correlation of the variables representing the measures of trait *i* measured by method *j* with the measure of trait *i'* measured by method *j'*. Heterotrait-monomethod triangles are grey-shaded; Heterotrait-heteromethod triangles are depicted with dashed lines. Convergent validities can be found in the cells on the main diagonals within the subtables (not belonging to any triangle).

Campbell and Fiske (1959) developed four criteria for evaluating the convergent and discriminant validity of measures within their MTMM framework:

1. The *correlations on the validity diagonal* $\left(\text{e.g., } r_{(T_2M_2,T_2M_1)}\right)$ depict the *convergent validity* of particular traits measured by different methods. These correlations should be significant and considerably high.
2. The *correlations on the validity diagonal (monotrait-heteromethod correlations)* should be higher than the correlations of the other variables of the same row or column in the particular *heterotrait-heteromethod block.* The measures of one trait by two different methods should be more strongly correlated (converge to a greater extent) than two different traits measured by the same two methods. Under these conditions, there is *discriminant validity*.
3. The *monotrait-heteromethod correlations* should be higher than the *heterotrait-monomethod correlations* $\left(\text{e.g., } r_{(T_2M_3,T_2M_1)} > r_{(T_2M_3,T_1M_3)}\right)$. This comparison also concerns the *discriminant validity*.
4. The correlations of variables should show the same patterns in all of the *heterotrait triangles* of both the *monomethod and heteromethod blocks*. This desideratum also concerns the *discriminant validity*. The associations of the different traits should be the same for all methods and all method combinations. Discriminant validity shall not depend on the set of methods used to measure the traits.

The guidelines presented by Campbell and Fiske (1959) still influence our modern understanding of validation. Marsh and Grayson (1995) give a good summary of the intention, impact, limitations, and consequences of the proposed guidelines:

> Campbell and Fiske (1959) were aware of most the limitations in their approach, specifically stating their guidelines should be viewed as “common sense desideratum” (p. 83). Their intent was to provide a systematic, *formative evaluation* of MTMM data at the level of the individual trait-method unit, qualified by the recognized limitations of their approach, not to provide a *summative evaluation* or global summaries of convergent validity, discriminant validity, and method effects. More generally, Campbell and Fiske had a heuristic intention to encourage researchers to consider the concepts of

> convergent validity, discriminant validity, and method effects; in this intention they were remarkably successful. (Marsh & Grayson, 1995, p. 180)

Modern statistical approaches as the Confirmatory Factor Analysis (CFA) in combination with structural equation modeling (SEM; especially Jöreskog, 1969, 1973) allow analyzing MTMM data with very sophisticated models (see e.g., Eid, 2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003; Eid, Lischetzke, & Nussbeck, 2006; Kenny, 1976, 1979; Kenny & Kashy, 1992; Marsh & Grayson, 1995; Saris & van Meurs, 1991; Widaman, 1985). All of these models allow for a separation of measurement error from latent scores, thus enabling researchers to analyze the latent MTMM matrix which is corrected for differences in the reliabilities of the measures. Therefore, more accurate estimations of the convergent and discriminant validity free from distortion by measurement error can be obtained.

2.2 The Latent Multitrait-Multimethod Matrix

Out of the great variety of different CFA-MTMM models, the Correlated Trait (CT) Model with rater-specific trait-variables (T_{jk}) comes closest to the original matrix proposed by Campbell and Fiske (1959). This model is depicted in Figure 2.1. In this model, a latent trait variable (T_{jk}) is introduced for all observed variables (Y_{ijk}) measuring the same trait (*j)* rated by the same rater (*k*). In Figure 2.1, there are two observed variables (*i*) for every combination of traits and raters. That is, the score on item Y_{212} indicates the rating on the 2nd indicator (i = 2) of the 1st trait (j = 1) for the 2nd (k = 2) rater. In order to have latent rater-specific trait-variables each rater has to provide at least two ratings. In this model the number of latent traits corresponds to the product of traits and raters (methods) (e.g., 3 x 2 = 6 latent traits). The model allows for a separation of trait-rater-specific effects from measurement error. The correlations of the rater-specific latent traits can be analyzed in the standard framework provided by Campbell and Fiske (1959). It is the analysis of a latent MTMM matrix.

The measurement equation of the CT model with method-specific trait variables is:

$$Y_{ijk} = \alpha_{ijk} + \lambda_{\mathrm{T}ijk} T_{jk} + E_{ijk} \, . \qquad (2.2.1)$$

where α_{ijk} is the intercept and $\lambda_{\mathrm{T}ijk}$ is the loading coefficient of indicator Y_{ijk} on Trait T_{jk}. E_{ijk} represents the measurement error variable.

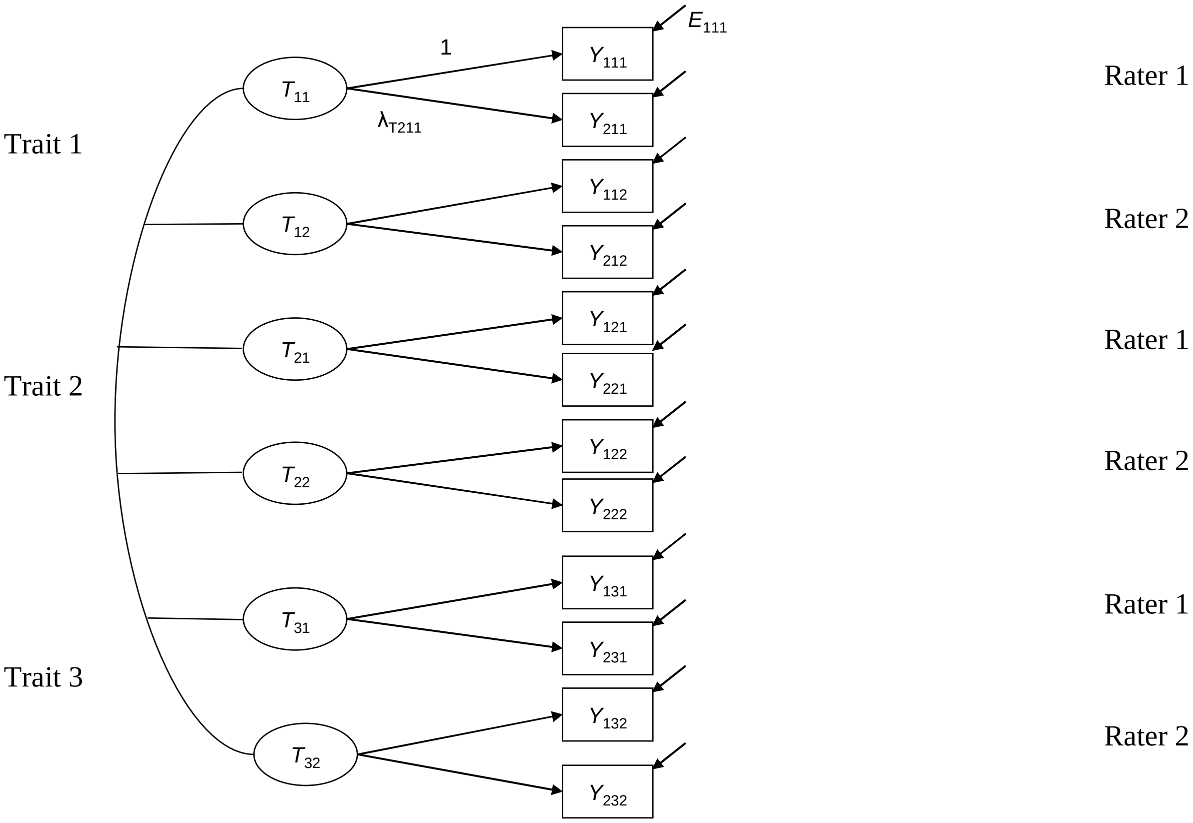

Figure 2.1. The CT-model for three constructs with method-specific trait variables. T_{jk}: trait variable; Y_{ijk}: observed variable; i: indicator; j: trait; k: rater; E_{ijk}: error variable (only depicted for the first indicator). Only the first two loading parameters are depicted $(\lambda_{\mathrm{T}111} = 1; \lambda_{\mathrm{T}112})$.

The six latent variables (presented in ovals in Figure 2.1) may be analyzed in the same way as the manifest variables presented in Table 2.1.1 Therefore, the convergent and discriminant validity can be determined on the latent level according to the criteria proposed by Campbell and Fiske (1959). A direct adoption of the statistical structure of the CT model to the analysis of categorical data is possible (see Hagenaars, 1990, 1993). However, no model for the analysis of latent rater agreement as well as the analysis of convergent and discriminant validity for categorical data has been formulated yet.

Therefore, the existing models of rater agreement will first be revised in order to adopt their structure on the latent level.

2.3 Manifest Rater Agreement Models

The analysis of rater agreement[3] has a long tradition in psychology as in the social sciences in general. Indices and models of rater agreement have mainly been proposed for the analysis of multivariate cross-classifications of non-ordered categorical (nominal) data. Non-ordered categorical variables are variables whose values only serve to identify categories without any quantitative meaning. Clinical disorders, for example, are often measured on a nominal scale. The assignment of “1” to “paranoid schizophrenia disorder” and “2” to “major depressive disorder” is equally admissible as the reverse. The assignment of numbers to the categories has no impact on the further analysis of the data, because nominal variables are not ordered in a specific manner. Nominal variables can be obtained by a wide array of different "ratings" such as self-ratings, peer ratings, medical, and psychological diagnoses (for an overview see e.g., Bakeman & Gnisci, 2006; Neyer, 2006). The assignment to categories requires that each and every observation is classified into one and only one category. The categories must be exhaustive and mutually exclusive.

Although categories have to be mutually exclusive, this does not imply that all raters provide the same score for the same object. This may be due to an inaccurate definition of the categories, differences in the amount and / or quality of information between raters, or to biased ratings by one or more raters. To analyze the convergence of different methods (the agreement between raters), nominal variables are usually presented in cross-classifications (cross tables), in which the rows and columns represent the different categories of the manifest variables measured by the different methods. The agreement between two or more raters can be determined relying on different indices of rater agreement. The analysis of rater agreement is not restricted to the case of nominal data but all indices and models presented in this dissertation may also serve to quantify the agreement (convergent validity) for scores of higher measurement levels (ordinal or interval level data).

[3] Large parts of this chapter have been published by Nussbeck (2006).

2.3.1 Rater Agreement Indices

The *proportion agreement index* (percentage agreement index) may be seen as an intuitive and useful first measure of agreement. It is simply the proportion of identical assignments of two raters. It is computed by:

$$p_o = \frac{\sum_{i=1}^{I} (n_{ii})}{\sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij})}, \tag{2.3.1}$$

where n_{ij} denotes the number of cases in cell *ij* of the table representing the cross-classification of the two ratings (*i*: rating of the 1st rater; *j*: rating of the 2nd rater), n_{ii} denotes the entries on the main diagonal (representing agreement, where *i* = *j*). Its range is from 0 to 1 with 1 indicating perfect agreement. Sometimes the proportion agreement index is referred to as percent agreement (Hartmann, 1977), interval-by-interval agreement (Hawkins & Dotson, 1975), exact agreement (Repp, Deitz, Boles, Deitz, & Repp, 1976), overall reliability (Hopkins & Hermann, 1977), total agreement (House, House, & Campbell, 1981), or point-by-point reliability (Kelly, 1977).

Unfortunately, as Suen and Ary (1989) have shown, the proportion agreement index is inflated by chance agreement and suffers from its dependency on the marginal distributions. Agreement on chance can simply be determined by multiplying the marginal proportions:

$$e_{ij} = \frac{n_{i+} n_{+j}}{N} \tag{2.3.2}$$

with e_{ij} depicting the expected proportion of cell *ij* given independent ratings and *N* is the sample size. “+” in the subscripts indicates the cells which have been collapsed. That is, the cells which have been added to yield a marginal frequency. Determining the expected cell frequencies using Eq. 2.3.2 for Table 2.3.1(b) shows that the observed cell frequencies exactly correspond to the expected frequencies under assumption of independence. There is no agreement beyond chance agreement for the two raters. However, the proportion agreement index is rather high $\left(p_O = \frac{0+445}{55+445} = .89 \right)$ implying considerable agreement.

Table 2.3.1

Two cross-classifications of two ratings (artificial data)

(a) *Data Set 1*

		Rater B		Marginal distribution of A
		1	2	n_{i+}
Rater A	1	40	15	55
	2	20	425	445
Marginal distribution of B	n_{+j}	60	440	500

(b) *Data Set 2*

		Rater B		Marginal distribution of A
		1	2	n_{i+}
Rater A	1	0	55	55
	2	0	445	445
Marginal distribution of B	n_{+j}	0	500	500

Note. n_{i+} represents the number of times rater A chooses categories 1 or 2, respectively. The corresponding frequencies for rater B are denoted by n_{+j}. These marginals are obtained by adding the cell counts of the corresponding row (or column, respectively).

Additionally, the proportion agreement index is not sensitive with respect to critical cases (hyperactive children, for example). This can best be illustrated by the data in Table 2.3.1(b). Assume, for example, that 55 pupils actually should be rated 1 (e.g. hyperactive, as does A correctly[4]) and 445 should be rated 2 (not hyperactive). As can be seen in Table 2.3.1(b), both raters agree 445 times diagnosing pupils as “2” while in the other 55 times,

[4] Assume that the “true” score for the pupils is known.

Rater A correctly judges “1” while B assesses the same pupils as “2”. The proportion agreement index yields a value of p_O = .89, which is quite similar to the value obtained from the data presented in Table 2.3.1(a) $\left(p_o = .93\right)$. However, both raters do not agree in even one critical case, whereas in the upper part of Table 2.3.1 both raters agree in 40 critical cases. The high agreement in Table 2.3.1(b) stems from the low prevalence of hyperactivity which is correctly reflected by the marginal distribution of Rater A. Because A correctly identifies hyperactive pupils, the proportion agreement index may lead to the improper conclusion that B does so as well. But this high level of agreement is completely due to the agreement between the two raters for cases belonging to category 2. Hence, the proportion agreement index severely suffers from its insensitivity to critical cases and its dependency on the distribution of the criterion (i.e., its prevalence). As the actual prevalence of behavior occurrence approaches unity or zero, the possibility increases that the proportion agreement index is inflated (Costello, 1973; Hartmann, 1977; Hopkins & Herman, 1977; Johnson & Bolstad, 1973; Mitchel, 1979). The closer the prevalence is to .50, the less likely the proportion agreement index is inflated (Suen & Ary, 1989). Unless one knows the marginals it is impossible to provide reasonable thresholds for the proportion agreement index.

The *occurrence and nonoccurrence agreement indices* can be used when the prevalence of a critical observation is very low or very high. The occurrence index (p_{occ}) should be used when the prevalence rate falls below .20. It is computed by:

$$p_{occ} = \frac{\text{occurrence agreements}}{\text{occurrence agreements + disagreements}}. \tag{2.3.3}$$

When the prevalence rate is higher than .80, the nonoccurrence agreement index (p_{non}) should be used (Kelly, 1977). The nonoccurrence agreement index is calculated by replacing the occurrence agreements by nonoccurrence agreements in Equation 2.3.3. The occurrence (or nonoccurrence, respectively) agreements reflect the number of times both raters agree on the occurrence (nonoccurrence) of the *critical* category and the disagreements reflect the times both raters disagree in general (on occurrence *and* nonoccurrence). Unfortunately, the occurrence (nonoccurrence) agreement index corrects for most of the agreement on chance, but not for the total agreement on chance since it still depends on the marginals (Suen & Ary, 1989). Another limitation is that no prior knowledge about the prevalence rates exists that would allow for a theoretically founded application of these indices.

The χ^2 - (chi-square) *value* as a measure of association can also be used to analyze rater agreement. Comparing the observed cell frequencies against their expected frequencies under the assumption of independence allows determining if some cells are more (less) often represented than expected by chance:

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}}, \qquad (2.3.4)$$

with $e_{ij} = \frac{n_{i+} n_{+j}}{N}$. n_{i+} and n_{+j} represent the marginals of row *i* and column *j*, respectively. High values indicate high associations of the ratings. The statistical significance of this measure of association can be determined by comparing the empirical value (Eq. 2.3.4) to the theoretically expected value given the degrees of freedom. The degrees of freedom of the corresponding χ^2-distribution can be determined by $df = \left(I - 1\right)^2$ for quadratic contingency tables. The higher the χ^2-value, the less the observed cell frequencies match the expected cell frequencies. One major drawback of the χ^2-statistic is its dependency on the sample size. Contingency tables with identical cell-proportions yield higher χ^2-values for those with larger samples.

The χ^2-value is not restricted to a special range of values. Its values are larger than zero but have no upper limit. To make its values more comparable, the corrected Contingency Coefficient $\mathrm{C}_{\mathrm{corr}}$ and Cramer’s V can be computed (see for example Liebetrau, 1983). Both coefficients transform the empirical χ^2-value to obtain values ranging from zero to one. In these transformations the empirical χ^2-value is compared to a maximal χ^2-value ($C_{\max}$). The transformed coefficients (C_{corr} or V) can be interpreted as a measure of association:

Contingency-Coefficient *C*:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}, \qquad (2.3.5)$$

Corrected Contingency Coefficient C_{corr}:

$$C_{corr} = \frac{C}{C_{\max}}, \qquad \text{with } C_{\max} = \sqrt{\frac{R-1}{R}} \text{ and } \mathrm{R} = \min(I, J). \tag{2.3.6}$$

Cramer's V:

$$V = \sqrt{\frac{\chi^2}{n(R-1)}}, \qquad \text{with } \mathrm{R} = \min(I, J). \tag{2.3.7}$$

Unfortunately $\mathrm{C}_{\mathrm{corr}}$ cannot reach 1 in nonquadratic contingency tables (where $I \neq J$), whereas V does. Both coefficients are hard to interpret because there is no standard for judging their magnitudes (Reynolds, 1977a, 1977b). Bishop, Fienberg, and Holland (1975) conclude that these coefficients should only be used for comparing several tables and may not be interpreted per se.

Coefficient kappa (κ; Cohen, 1960) is a flexible index that is, applicable to dichotomous or polytomous variables involving two or more observers. κ is computed by:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \tag{2.3.8}$$

where P_o represents the observed proportion of identical ratings $\left(P_o = \sum_{i=1}^{I} p_{ii}\right)$ and P_e the expected proportion of agreement by arbitrary ratings $\left(P_e = \sum_{i=1}^{I} p_{i+} p_{+i}\right)$, p_{ij} denotes the proportion of observations within each cell $\left(p_{ij} = \frac{n_{ij}}{N}\right)$, whereas I denotes the number of categories.

κ ranges from –1.00 to +1.00, whereby a positive κ indicates that the observers agree more frequently than expected by chance, zero indicates that both raters agree on the same level as expected by chance and a negative value indicates that both raters agree less often than expected by chance. A negative κ provides a strong hint that raters do not use all categories in the appropriate way. As a rule of thumb, a κ of .60 can be regarded as the minimal acceptable level of agreement (Gelfland & Hartmann, 1975) whereas a κ of .80 is an indication of high agreement (Landis & Koch, 1977).

2.3.2 Advantages and Limitations of Rater Agreement Indices

In general, associations between variables or methods can be detected by the χ^2-value as a measure of association. This value can also be compared to its theoretical distribution yielding the χ^2-test. This test is principally conducted on the basis of the null hypothesis that all variables are independent from each other. The χ^2-value provides information on whether the data differ significantly from the expected cell frequencies. Information about the strength of association can be obtained by the corrected Contingency Coefficient and Cramer’s *V*.

The special case of rater agreement can be analyzed by several methods. As pointed out, many of them are afflicted by specific problems. The most promising approach seems to be the κ-coefficient, a method that is a chance-corrected version of proportion agreement. Suen, Ary, and Ary (1986) demonstrated the mathematical relationship between κ and proportion agreement and also provided conversion procedures from one index to the other.

Many authors suggest κ to be the most preferable agreement index because it corrects for chance agreement, is related to percentage (proportion) agreement, and is comparable between studies (see Suen & Ary, 1989) while others criticize it as not comparable between studies (Cicchetti & Feinstein, 1990; Feinstein & Cicchetti, 1990; Thompson & Walter, 1988a, 1988b; Uebersax, 1987). Indeed, κ can be used to test whether ratings agree to a greater extent than expected by chance. Yet, there is still concern about using κ as a measure of agreement because it is only chance-corrected for the assumption of independent ratings, an assumption which is implicitly made but legitimated by no means (it is assumed that chance agreement is based on the independence model). Uebersax (1987) demonstrated how differences in the accuracy with which positive and negative cases can be detected (i.e., differences in the mathematical characteristics of the particular decision-making process) affect the value of κ. Therefore, it is not useful to compare κ across studies. Moreover, this problem increases when there are different base rates. In general, if the sample consists of cases which belong to an easily identifiable category, a higher κ is obtained, although the diagnostic accuracy remained the same compared to a sample consisting of less easily identifiable cases.

Diagnosability curves representing the degree to which diagnosticians are able to accurately judge subjects with respect to the subjects' true status may actually differ so much that κ-values obtained for the same symptom (criterion) with similar base rates cannot be compared across studies. Unless there is an explicit model of rater decision making, it remains unclear how chance affects decisions of actual raters and how one might correct for it (Uebersax, 1987).

Increasing the number of categories is no problem for the different rater agreement indices. However, when the number of methods (observers) increases, the application of the general agreement indices becomes more complicated. In this case, κ should be determined for each rater pair, and the median value should be taken as the overall value (Conger, 1980; Fleiss, 1971). For example, Fleiss (1971) developed modifications of κ to determine rater agreement when objects are rated by the same number of raters to compute agreement with regard to a particular object, and to estimate agreement within a particular category.

A high level of agreement between raters does not guarantee an individually correct diagnosis; yet, disagreement between raters often indicates a lack of diagnostic accuracy (Uebersax & Grove, 1990). The association between variables and the extent to which methods or raters agree depend on two major criteria. First, it is important that both raters can well distinguish between any pair of categories. *Distinguishability* between two categories increases if the ratio of concordant ratings to discordant ratings of different observers increases. The second criterion is the lack of *bias* (Agresti, 1992). According to Agresti's definition, the amount of bias depends on the comparison of the marginal distributions: If raters use the response categories with the same frequency, their marginal distributions are homogeneous, indicating that none of the raters prefers a particular category compared to the other raters. However, homogeneous marginal distributions do not imply that all raters judge the subjects correctly compared to the subjects' true status, but they show that they use the response categories in a similar way. If all raters distinguish between categories in the same way and their marginal distributions are similar, subjects will be more congruently assigned to the categories of a variable, thus providing hints that observers define the categories in a similar way.

2.3.3 Rater Agreement Models

All general agreement indices described so far fail to provide more detailed information about various types and sources of agreement and disagreement. However, this kind of information can be obtained by modeling associations between variables using *log-linear* models. For special cases of association, effect sizes (as the χ^2-value or Cramer's *V*) can be estimated representing the degree of association between variables. Conditional probabilities of receiving a particular response by an observer given the responses of other observers can be computed. Finally residuals can be determined that compare the frequencies with which certain types of agreement and disagreement occur compared to what would be expected with some predicted pattern (Agresti, 1990, 1992).

All log-linear models for the common distributions of two variables are restricted models of the *saturated* log-linear model:

$$e_{ij} = \eta \tau_i^A \tau_j^B \tau_{ij}^{AB}, \tag{2.3.9}$$

where the expected cell frequency (e_{ij}; with i = 1,…I and j = 1,…, J denoting the categories) is computed by the product of the overall effect (η), two one-variable effects (τ_i^A, τ_j^B), and the two-variable effect (τ_{ij}^{AB}). In the saturated (population) model, the model parameters can be determined by simply comparing frequencies and mean frequencies of different cells of the joint distribution of different variables. The estimation of the parameters for other models has to be done using Maximum-Likelihood (ML) procedures. Table 2.3.2 depicts the joint distribution of two variables (their cross-classification). The extension to more than two variables is straightforward.

The overall effect (η) represents the geometric mean of all cell frequencies and is, thus a mere reflection of the sample size (Hagenaars, 1993). It can be determined by:

$$\eta = \sqrt[IJ]{\prod_{i=1}^{I} \prod_{j=1}^{J} e_{ij}}\ . \tag{2.3.10}$$

The one-variable effects $\left(\tau_i^A, \text{ and } \tau_j^B\right)$ reflect deviations of the geometric mean of all cells belonging to the *i*th (respectively, *j*th) category of a variable. They can be estimated by:

$$\tau_i^A = \frac{\sqrt[J]{\prod_{j=1}^{J} n_{ij}}}{\eta},$$

and (2.3.11)

$$\tau_j^B = \frac{\sqrt[I]{\prod_{i=1}^{I} n_{ij}}}{\eta}.$$

Table 2.3.2

Cross-classification of two variables

		Variable *B*			
		1	…*j*…	*J*	n_{i+}
Variable *A*	1	n_{11}	…	n_{1J}	n_{1+}
	…*i*…	…	…	…	$\dots n_{i+} \dots$
	I	n_{I1}	…	n_{IJ}	n_{I+}
	n_{+j}	n_{+1}	$\dots n_{+j} \dots$	n_{+J}	*N*

Note. …*i*… and …*j*… indicate specific categories of the finite number of categories for *I* and *J*.

In the saturated model, all cell frequencies are exactly reproduced. Therefore, the one-variable effects reflect the odds comparing a particular marginal to the overall effect.

The one-variable effect gives first insight into rater-bias (or method bias *MB*; with respect to the other rating[5]). Ratings are biased with respect to each other to the degree their marginal distributions differ from each other (Agresti, 1992):

[5] I will refer to these rater-specific effects as method bias to be in line with the existing literature (i.e., Agresti, 1992).

$$MB_{(A/B)} = \frac{\pi_i^A}{\pi_i^B}, \qquad (2.3.12)$$

with *i* indicating the identical category of raters A and B. *MB* in Equation 2.3.12 is the rater-effect of Rater A for category *i* compared to Rater B (*A* / *B*). A value greater than 1 indicates a higher proportion (a value smaller than 1 a smaller proportion) of this category for rater A than for rater B. This kind of rater effect can be determined in all following models (relying on the expected frequencies or proportions). For the saturated model, the rater-bias can directly be computed relying on the ratio of the log-linear parameters $MB_{(A/B)} = \frac{\tau_i^A}{\tau_i^B}$. It is the degree to which A or B overestimates (underestimates) the prevalence of a particular category with respect to the other rater. It is especially meaningful to calculate this index if one of the raters provides “better” ratings than the other. That is, if one rater can be seen as a gold standard (like a reference method, a well established method) it is meaningful to compare the other rater against this gold-standard rater.

Finally, the two-variable effect $\left(\tau_{ij}^{AB}\right)$ depicts the deviation of a particular cell from its expected value given the overall and one-variable effects. It corresponds to the odds of the actual observed cell frequency with respect to the expectation given the overall effect and the two odds depicting the deviation of the corresponding row $\left(\tau_i^A\right)$ and column $\left(\tau_j^B\right)$ from the overall geometric mean:

$$\tau_{ij}^{AB} = \frac{e_{ij}}{\eta\tau_i^A\tau_j^B}. \qquad (2.3.13)$$

The saturated model exactly reproduces the observed cell frequencies; it does not impose any restriction on the expected frequencies and therefore does not contain testable consequences.

A useful first analysis of agreement can be done by testing the *independence model*. The independence model assumes that there is no association between both raters[6]. Thus,

[6] The log–linear models for rater agreement are generally introduced for the case of two raters but can be extended to more than two raters.

the two-variable effect-parameters $\left(\tau_{ij}^{AB}\right)$ are set to 1. The model equation for the independence model appears as:

$$e_{ij} = \eta \tau_i^A \tau_j^B . \tag{2.3.14}$$

In this model, only the one-variable effects are implemented which means that the marginal distributions of both variables are reproduced. If these one-variable effects are equal to each other $\left(\tau_i^A = \tau_j^B, \text{ for } i = j\right)$, both variables' marginal distributions are homogeneous. Homogeneous marginal distributions imply that both raters choose each category with the same frequency; accordingly, no rater prefers any category to a greater extent than the other, which means that no rating is biased (with respect to the other rater; Agresti, 1992). This type of model only rarely fits empirical data because, in general, different measures of a construct are related to a certain degree representing the convergent validity.

Useful information provided by the independence model stems from the analysis of its adjusted cell residuals. Adjusted cell residuals compare observed with expected cell frequencies (see Agresti, 1992):

$$r_{ij} = \frac{n_{ij} - e_{ij}}{\sqrt{e_{ij}\left(1 - \frac{n_{i+}}{n_{++}}\right)\left(1 - \frac{n_{+j}}{n_{++}}\right)}} . \tag{2.3.15}$$

A useful extension of the independence model is the *quasi-independence model*. In this model, a new parameter is introduced. This parameter is only implemented for cells on the main diagonal which represent agreement between methods:

$$e_{ij} = \eta \tau_i^A \tau_j^B \left(\tau_{ij}^{AB}\right)^I, \text{ with } I = \begin{cases} 1, \text{ if } i = j \\ 0, \text{ if } i \neq j \end{cases} . \tag{2.3.16}$$

In contrast to the independence model, the *quasi-independence model I* allows for higher cell frequencies in cells on the main diagonal, but no overrepresentation in any other cell. For cells indicating disagreement, the independence model holds. As a result of the newly introduced parameters $\left(\tau_{ij}^{AB}\right)^{I}$, the estimated cell frequencies on the main diagonal indicating agreement exactly match the empirical cell frequencies. The parameters $\left(\tau_{ij}^{AB}\right)^{I}$ can be used to compare the probability of receiving a particular response by one method given the rating of the other method (see Agresti, 1992). The probability to find an observation in a particular cell on the main diagonal is $\left(\tau_{ij}^{AB}\right)^{I}$ times larger than expected by chance (represented by an independence model). Sometimes the parameter $\left(\tau_{ij}^{AB}\right)^{I}$ is also presented as $\left(\tau_{ii}^{AB}\right)^{I}$ indicating that A and B both choose the same category *i*. Bias (with respect to the other rater) can be examined as in the independence model.

If all parameters $\left(\tau_{ij}^{AB}\right)^{I}$ are equal to each other, all expected cell frequencies on the main diagonal differ from chance agreement to the same degree. Hence, a simpler model holds which assumes $\left(\tau_{ij}^{AB}\right)^{I}$ to be constant:

$$e_{ij} = \eta\tau_i^A\tau_j^B\left(\tau^{AB}\right)^{I}\text{, with } I = \begin{cases}1, \text{ if } i = j\\ 0, \text{ if } i \neq j\end{cases}. \qquad (2.3.17)$$

In this *quasi-independence II model*, the sum of the expected cell frequencies on the main diagonal is exactly equal to the sum of the observed frequencies whereas single expected cell frequencies on the main diagonal may differ slightly. The difference between both models is that in the latter, the degree of agreement between both methods is the same for all categories under consideration, whereas in the first, agreement between methods may differ from category to category.

Table 2.3.4 presents the cells of a cross-classification of two observed variables' proportions $\left(\hat{\pi}_{ij}^{AB} = \frac{e_{ij}}{N}\right)$ for the quasi-independence I model. The cells present the proportions and the underlying log-linear model parameters. All proportions for cells besides the main diagonal only depend on one-variable effects implying independence. The cells on the main diagonal additionally depend on two-variable effects.

The fitted cell proportions in estimations of this model are the cells on the main diagonal. That is, their expected proportions equal the observed proportions. All other expected proportions may deviate from the observed proportions. Schuster and Smith (2006) showed how the quasi-independence model can be represented as a mixture distribution model separating ambiguous from obvious cases. Their approach is to split a population for which the quasi-independence model holds into two sub-populations. For the first sub-population (the ambiguous cases) the independence model holds (see Table 2.3.5)—that is, all raters independently rate individuals of the population—for the second sub-population (obvious cases) a one variable model holds (see Table 2.3.6)—that is, all ratings depend perfectly from each other all raters rate every individual perfectly congruently. In the latter subpopulation, all individuals are cross-classified on the main diagonal and hence, one variable is sufficient to describe the relationship. Recall, that raters may agree upon ambiguous cases but only due to chance agreement. For the subpopulation of obvious cases the one-variable models implies that there is perfect agreement.

Schuster and Smith (2006) related the quasi-independence II parameter for cells on the main diagonal to κ. However, the meaning of the log-linear parameters has not been described yet. Tables 2.3.5 and 2.3.6 show how the different log-linear effects influence the cell proportions. Box 2.3.1 gives an overview on their statistical meaning. The log-linear parameters of the quasi-independence models cannot be easily linked to proportions, odds, or odds ratios. Drawing a parallel to Hagenaars (1993): In order to understand the implications of the model, the model should be estimated and its expected proportions should be interpreted rather than its parameters should be inspected.

The same rationale as presented for the quasi-independence I model presented in Tables 2.3.5 to 2.3.6 and Box 2.3.1 also accounts for the quasi-independence II model. The only difference is that the two-variable log-linear parameters are restricted to be constant. In both models, the rater-bias coefficient (*MB*) may be used to determine the influences of rater-specific effects.

Table 2.3.4

Parameters in the quasi-independence I model

		Variable *B*			
		b = 1	*b* = 2	*b* = 3	
Variable *A*	*a* = 1	$\pi_{11}^{AB}=\frac{\tau_1^A\tau_1^B\tau_{11}^{AB}}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\left(\tau_j^B\right)+\tau_{11}^{AB}}$	$\pi_{12}^{AB}=\frac{\tau_1^A\tau_2^B}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\tau_j^B}$	$\pi_{13}^{AB}=\frac{\tau_1^A\tau_3^B}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\tau_j^B}$	π_{1+}^{AB}
	a = 2	$\pi_{21}^{AB}=\frac{\tau_2^A\tau_1^B}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\tau_j^B}$	$\pi_{22}^{AB}=\frac{\tau_2^A\tau_2^B\tau_{22}^{AB}}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\left(\tau_j^B\right)+\tau_{22}^{AB}}$	$\pi_{23}^{AB}=\frac{\tau_2^A\tau_3^B}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\tau_j^B}$	π_{2+}^{AB}
	a = 3	$\pi_{31}^{AB}=\frac{\tau_3^A\tau_1^B}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\tau_j^B}$	$\pi_{32}^{AB}=\frac{\tau_3^A\tau_2^B}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\tau_j^B}$	$\pi_{33}^{AB}=\frac{\tau_3^A\tau_3^B\tau_{33}^{AB}}{\sum_{i=1}^{3}\left(\tau_i^A\right)\sum_{j=1}^{3}\left(\tau_j^B\right)+\tau_{33}^{AB}}$	π_{3+}^{AB}
		π_{+1}^{AB}	π_{+2}^{AB}	π_{+3}^{AB}	

Table 2.3.5

Independence sub-table in the quasi-independence I model (ambiguous cases in Schuster & Smith, 2006)

	b = 1	b = 2	b = 3	
a = 1	$\pi^{\circ}_{11}=\frac{\tau^{A}_{1}\tau^{B}_{1}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi^{\circ}_{12}=\frac{\tau^{A}_{1}\tau^{B}_{2}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi^{\circ}_{13}=\frac{\tau^{A}_{1}\tau^{B}_{3}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi_{1+}-\frac{\tau^{AB}_{11}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\left(\tau^{B}_{b}\right)+\tau^{AB}_{11}}$
a = 2	$\pi^{\circ}_{21}=\frac{\tau^{A}_{2}\tau^{B}_{1}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi^{\circ}_{22}=\frac{\tau^{A}_{2}\tau^{B}_{2}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi^{\circ}_{23}=\frac{\tau^{A}_{2}\tau^{B}_{3}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi_{2+}-\frac{\tau^{AB}_{22}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\left(\tau^{B}_{b}\right)+\tau^{AB}_{22}}$
a = 3	$\pi^{\circ}_{31}=\frac{\tau^{A}_{3}\tau^{B}_{1}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi^{\circ}_{32}=\frac{\tau^{A}_{3}\tau^{B}_{2}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi^{\circ}_{33}=\frac{\tau^{A}_{3}\tau^{B}_{3}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\tau^{B}_{b}}$	$\pi_{1+}-\frac{\tau^{AB}_{33}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\left(\tau^{B}_{b}\right)+\tau^{AB}_{33}}$
	$\pi_{+1}-\frac{\tau^{AB}_{11}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\left(\tau^{B}_{b}\right)+\tau^{AB}_{11}}$	$\pi_{+2}-\frac{\tau^{AB}_{22}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\left(\tau^{B}_{b}\right)+\tau^{AB}_{22}}$	$\pi_{+3}-\frac{\tau^{AB}_{33}}{\sum_{a=1}^{3}\left(\tau^{A}_{a}\right)\sum_{b=1}^{3}\left(\tau^{B}_{b}\right)+\tau^{AB}_{33}}$	

Note. The probabilities presented in this table do not correspond directly to the probabilities in the text since the complete latent table is split into two parts. For reasons of readability the superscripts *AB* are not depicted for the proportions $\left(\pi\right)$.

Table 2.3.6

Agreement (Reliability) sub-table in the quasi-independence I model (obvious cases in Schuster & Smith, 2006)

	b = 1	b = 2	b = 3	
a = 1	$\pi^*_{11} = \frac{\tau^{AB}_{11}}{\sum_{i=1}^{3}\tau^{AB}_{ii}}$			π^*_{11}
a = 2		$\pi^*_{22} = \frac{\tau^{AB}_{22}}{\sum_{i=1}^{3}\tau^{AB}_{ii}}$		π^*_{22}
a = 3			$\pi^*_{33} = \frac{\tau^{AB}_{33}}{\sum_{i=1}^{3}\tau^{AB}_{ii}}$	π^*_{33}
	π^*_{11}	π^*_{22}	π^*_{33}	

Note. The probabilities presented in this table do not correspond directly to the probabilities in the text since the complete latent table is split into two parts. For reasons of readability the superscripts *AB* are not depicted for the proportions (π).

Box 2.3.1

For the independence model it is known (see e.g., Hagenaars, 1993):

$$\tau^A_i = \frac{\pi^A_i}{\sqrt[3]{\prod_{a=1}^{3}\pi^A_a}}, \text{ and } \tau^B_j = \frac{\pi^B_j}{\sqrt[3]{\prod_{b=1}^{3}\pi^B_b}}, \qquad (2.3.18)$$

with *a* indicating the categories of A in the independence table and *b* indicating the categories of B in the independence table.

Therefore:

$$\tau^A_i = \frac{\pi^{AB}_{i+} - \dfrac{\tau^{AB}_{ii}}{\sum_{a=1}^{3}\tau^A_a\sum_{b=1}^{3}\tau^B_b + \tau^{AB}_{ii}}}{\sqrt[3]{\prod_{a=1}^{3}\pi^A_a}}, \text{ and } \tau^B_j = \frac{\pi^{AB}_{+j} - \dfrac{\tau^{AB}_{jj}}{\sum_{a=1}^{3}\tau^A_a\sum_{b=1}^{3}\tau^B_b + \tau^{AB}_{jj}}}{\sqrt[3]{\prod_{b=1}^{3}\pi^B_b}} \qquad (2.3.19)$$

showing that the one-variable parameters do not exclusively relate to the marginal proportions. Combining Tables 2.3.5 and 2.3.6 yields: $\pi^{\circ}_{ii} + \pi^*_{ii} = \pi_{ii}$, for the total table.

Replacing:

$$\pi_{ii}^{\circ} + \pi_{ii}^{*} = \pi_{ii} = \frac{\tau_i^A \tau_i^B \sum_{a=1}^{3} \tau_{aa}^{AB} + \tau_{ii}^{AB} \sum_{a=1}^{3} \tau_a^A \sum_{b=1}^{3} \tau_b^B}{\sum_{a=1}^{3} \tau_a^A \sum_{b=1}^{3} \tau_b^B \sum_{a=1}^{3} \tau_{aa}^{AB}}$$

$$\Leftrightarrow \frac{\pi_{ii}\left(\sum_{a=1}^{3} \tau_a^A \sum_{b=1}^{3} \tau_b^B \sum_{a=1}^{3} \tau_{aa}^{AB}\right) - \tau_i^A \tau_i^B \sum_{a=1}^{3} \tau_{aa}^{AB}}{\sum_{a=1}^{3} \tau_a^A \sum_{b=1}^{3} \tau_b^B \sum_{a=1}^{3} \tau_{aa}^{AB}} = \tau_{ii}^{AB} \sum_{a=1}^{3} \tau_a^A \sum_{b=1}^{3} \tau_b^B , \qquad (2.3.20)$$

$$\Leftrightarrow \frac{\pi_{ii}\left(\sum_{a=1}^{3} \tau_a^A \sum_{b=1}^{3} \tau_b^B\right) - \tau_i^A \tau_i^B}{\left(\sum_{a=1}^{3} \tau_a^A \sum_{b=1}^{3} \tau_b^B\right)^2} = \tau_{ii}^{AB}$$

identifies the statistical meaning of the two-variable effect. This parameter cannot easily be related to a category proportion.

Log-linear models of agreement can also satisfy the property of *quasi-symmetry* (Darroch & McCloud, 1986). Because there is no objectively precise definition of how to classify an observation into the different categories for most cases in the social sciences, the discrepancies between classifications by different methods are attributable to measurement error and to different perceptions or interpretations of what a category definition means. “The correct category for an object exists partially in the eye of the beholder” (Darroch & McCloud, 1986, p. 376). On the other hand, there are signals sent out by each object which partially conform to each of the categories to a certain degree. These signals are assumed to differ between objects. Thus, the classification of an object into a particular category depends on the signals sent out by the object and the rater-specific category definition. If raters perceive these signals but may confuse their meaning (differ in their category definitions) a symmetric pattern of disagreement should occur:

$$e_{ij} = \eta \tau_i^A \tau_j^B \tau_{ij}^{AB}, \text{ with } \tau_{ij}^{AB} = \tau_{ji}^{AB} \text{ for all } i \text{ and } j . \qquad (2.3.21)$$

Hence, this model does not only address agreement between raters and indicates rater bias with respect to the marginal distributions, but additionally provides some information about rater-specific effects (rater bias; Agresti, 1992). This model is called the *quasi-symmetry model* because the expected cell frequency to receive a particular response by the first rater (say category *i*) and a particular response by the second rater (say category *j*) differs by the same ratio $\left(\tau_{ij}\right)$ from the expected cell frequency given only the one-variable effects as the contrary combination [*j i*][7]. In other words, associations between both raters are "mirrored" around the main diagonal.

Therefore, information about rater-specific effects can be obtained inspecting the *MB*-coefficient. If this coefficient differs from 1, the observers have different classification probabilities for the objects which means that they do not use the categories in the same manner. Additionally, inspecting the two-variable effects yields information to which degree particular category combinations are more or less frequent. That is, if the two raters confound categories in the same way. Assume that rater A correctly rates all individuals (knowing the true category of the individuals), the two-variable effects then indicate to which degree B agrees with A or if B systematically confounds categories $\left(\tau_{ij}^{AB}; i \neq j\right)$. Yet, it could also be the case that B correctly rates all individuals and A systematically confounds categories $\left(\tau_{ij}^{AB}; i \neq j\right)$—restricting $\left(\tau_{ij}^{AB} = \tau_{ji}^{AB}\right)$[8] thus yields identical systematic interactions and, thus, raters are interchangeable with respect to their confounding of categories.

If the one-variable effects do not differ between raters the more restrictive assumptions of the *symmetry model* hold. Formally, the symmetry model appears to be quite similar to the quasi-symmetry model:

$$e_{ij} = \eta \tau_i^A \tau_j^B \tau_{ij}^{AB}, \text{ with } \tau_{ij}^{AB} = \tau_{ji}^{AB} \text{ for all } i \text{ and } j, \text{ and } \tau_i^A = \tau_j^B \text{ for } i = j. \qquad (2.3.22)$$

In contrast to the quasi-symmetry model, the one-variable effects are set equal to each other. Thus, the marginal distributions of both variables have to be identical meaning that both raters agree on the prevalence of the categories (all *MB* = 1). In this model, the expected cell frequency of contrary combinations of categories is the same. Thus, the raters can be conceived interchangeable (Agresti, 1992).

[7] [a…. z] will be used throughout this thesis to indicate observed or expected patterns of categorical variables. That is, [*j i*] indicates that the rater A chooses category "*j*" and rater B chooses category "*i*".

[8] Interchanging the indices *i* and *j* signifies that the numbers of the categories are also interchanged from the left hand side of the equation to the right hand side of the equation.

2.3.4 Advantages and Limitations of Manifest Rater Agreement Models

The rater agreement models differ with respect to their implications for the kind of agreement (category-specific or general) and the interchangeability of raters. Compared to the quasi-independence models the quasi-symmetry as well as the symmetry model yield the benefit that observer differences *and* category distinguishability can be examined in detail (Darroch & McCloud, 1986) because both agreement *and* disagreement have to be modeled. If the quasi-symmetry model holds we can presume that raters produce the same amount of under- or overrepresentation for given combinations of categories and are thus interchangeable to their confounding of categories. Moreover, if the symmetry model holds, both raters are completely interchangeable (Agresti, 1992). A better fitting symmetry model compared to the quasi-symmetry model indicates a stronger association between ratings and interchangeability of raters. Interchangeable raters are also referred to as homogenous raters (see e.g., Schuster, 2002; Schuster & Smith, 2002, 2006; Zwick, 1988). These models allow for a test if the assumption to have interchangeable raters as a result of the research design is met.

As has been shown, there are different ways to measure agreement and disagreement by general agreement indices. In general, associations can be detected by the χ^2-test and, as a special case of association, rater agreement may be detected by coefficient kappa (κ). Model-based analysis of associations yields additional and more precise information than that provided by general association methods. Log-linear models allow testing of the goodness-of-fit (not only against independence as the χ^2-test). They provide model-implied fitted cell probabilities and enable researchers to make predictions of classifications under certain conditions such as receiving a particular response by an observer given the responses of other observers, receiving a response knowing the correct status of an observation, or assessing the latent status of an observation given ratings by several observers (Agresti, 1990, 1992; Bishop et al., 1975; Goodman, 1978; Haberman, 1978, 1979; Hagenaars, 1990). Thus, first analyses of rater agreement—as a special variant of convergence between multiple methods—can be conducted by overall agreement indices. These indices reveal if the raters tend to choose identical response categories. However, these indices only consider absolute agreement between raters (identical categories). More detailed information about the joint distribution of ratings is only available by use of log-linear models.

Log-linear models allow for a more fine graded analysis of rater agreement and disagreement. In this framework, categories can differ with respect to their agreement and disagreement rates. These rates may differ from one category to the other (see Table 1.1.1). Each score of one variable may have high co-occurrence with any other score of another variable allowing for a deeper understanding of the relations between variables. Log-linear models, for example, may reveal that the middle category of one variable co-occurs more frequently than expected based on the assumption of independence with the middle category of another variable. All other categories may not co-occur more or less frequently than expected by chance (their log-linear parameters do not differ significantly from 1).

All indices and models presented so far suffer from one major limitation. They do not allow for the analysis of more than one construct measured by one indicator per rater. Therefore, all information retrieved is specific to the combination of the trait (construct), the raters, and the indicator. Assuming that rater agreement depends on the items administered (some items are hard to judge, e.g., having self-doubts), the construct (some may be more easily detected, e.g., sociability; Funder , 1995), and the raters (peers may be better raters than acquaintances), it is necessary to extent the existing models to more indicators, more traits, and more raters.

Extending rater agreement models to models with multiple indicators per construct would allow for identifying underlying latent categories (so called classes, types, or statuses) which cause the different response patterns (observed scores on the multiple indicators). Many statuses of individuals can *not* be directly observed (e.g., psychiatric syndromes and disorders) but have to be deduced relying on multiple observations (which themselves may be classifications of overt behavior). If, for example, a researcher is interested in the adequacy of psychiatric diagnoses of different raters relying on the DSM-IV TR (American Psychiatric Association, 2004) it may be worthwhile not only to examine the final classification but to inspect the ratings of the single check-list categories. This inspection can reveal if a) all raters agree with respect to the check-list categories, b) if they come to the same conclusions about the status of the patient, c) if all categories are weighted to the same degree across raters to produce the final diagnoses, and d) if the categories of the observed variables reliably describe the latent variables. Latent (as manifest) rater agreement models could allow for a detailed analysis on which categories different raters agree, which categories indicating disagreement are only rarely chosen, and which categories indicating disagreement are chosen to a greater extent than expected for independent ratings. Integrating additional constructs (multiple traits) would allow for an analysis if there is higher or lower agreement for particular constructs and how the different categories of the different latent variables co-occur (free from measurement error) yielding information about discriminant validity.

3 Research Question

Determining the reliability and validity of different ratings is very important in many areas of psychology as pointed out in Section 1. Large MTMM studies yield information about convergent and discriminant validity of different scales. These analyses are mostly done for metric observed variables (for an overview see Eid, Lischetzke, & Nussbeck, 2006) or in some cases also for variables with ordered categorical response categories (see e.g., Nussbeck, Eid, & Lischetzke, 2006). The aim of this dissertation is to adopt the logic of MTMM models to the case of categorical data in general.

As pointed out in the previous sections, rater agreement models can be used to analyze agreement (convergent validity) and disagreement for observed manifest variables. However, we lack models that allow for determining the reliability of the manifest ratings and that allow for an inspection of agreement and disagreement free of influences due to measurement error. These models shall be developed in a first step. The parameters and / or (conditional) probabilities of the models will be linked to each other providing additional information about category-specific agreement rates, rater bias, and distinguishability of the latent categories. An empirical application will illustrate the meaning of the model parameters.

In a second step, a Multitrait-Multirater (MTMR) model for categorical data will be defined. This model will be based upon the latent rater agreement models of the first step enlarging their perspective to the analysis of discriminant validity. Additionally, the influence of particular latent statuses on agreement and / or disagreement may be analyzed. An empirical application will serve to illustrate the model.

The development of the latent rater agreement models and the MTMR models for categorical data is organized as follows:

- In a first step (4.1), the log-linear model with one latent variable will be introduced. This model serves to define the measurement structure of the latent variable. The measurement structure remains the same across all models and will therefore be presented in detail.
- In a second step (4.2), the model will be extended to a two latent variable model (see e.g., Hagenaars, 1990, 1993; Langeheine, 1988). This model is well introduced and serves as a basis for the introduction of latent rater agreement models. The meaning of the different model parameters will be explained.

- In a third step (5), the latent rater agreement models will be defined. Based on the first and second step, the different manifest rater agreement models will be adopted to the latent level. The different implications of these models will be explained in detail. These models allow for identifying very interesting pieces of information with respect to the agreement and disagreement of raters

  I will show how these models reveal i) if *raters agree* with each other, ii) if *raters agree in a general way* (irrespectively of the category under consideration) or if rater agreement is category specific, iii) if *disagreement* is less frequently expected than predicted by chance and if so, if this is the case in a general way or if there are some categories raters may better *distinguish* than others, iv) if some *disagreement combinations* are more often expected than predicted by chance implying a kind of confusion or lack of category-specific convergent validity, and v) if raters are *biased* with respect to the other rater.

  The latent rater agreement models will be defined for the case of structurally different and interchangeable raters. Most emphasis is paid to the interpretation of the model parameters and their theoretically meaningful deduction. Empirical applications serve to illustrate these models.

- In a fourth step (6), the latent rater agreement models will be extended to Multitrait-Multirater (MTMR) models. Integrating an additional rater agreement model into the saturated and symmetry latent rater agreement models described in the third step enlarges the agreement and disagreement analysis allowing for the analysis of discriminant validity.

  These models allow for determining if raters can use different pieces of information in a more specific (indicative) way for a given trait knowing the status of the other trait. Extraverted individuals may be rated more congruently on their emotions than others for example.

  Additionally, these models allow for the detailed analysis of overall agreement rates. That is, they allow for determining if raters agree on one construct with a higher probability if they also agree on the other construct. This effect reflects if there are good targets who can be congruently rated on both constructs. In the same vain, these models allow determining if specific disagreement combinations are more often expected yielding some information about which categories may be easily confounded by different raters.

- Finally (7), the models will be discussed with respect to their implications on agreement and disagreement, convergent and discriminant validity, rater-specific effects, and their relation to the theoretical framework of the rater accuracy model (RAM; Funder, 1995).

4 Latent Variable Models for Categorical Data

In this section, the framework of log-linear models with latent variables will be introduced (e.g., Goodman, 1974a, 1974b; Habermann, 1979; Hagenaars, 1990, 1993; McCutcheon, 1987; Vermunt, 1997b). In section 4.1, the most basic model for one construct measured by several items administered to one rater will be introduced. An empirical application serves to illustrate the meaning of the model parameters. In section 4.2, an additional latent variable will be introduced (see e.g., Hagenaars, 1990, 1993). The model will be defined and the meaning of the log-linear model parameters will be explained. An empirical application serves to illustrate the meaning of the model parameters.

4.1 Latent Variable Models for Categorical Data

Latent variable models for non-ordered categorical data have been developed during the last four decades. The two main approaches are the latent class (LCA) models and log-linear models with latent variables. LCA models have mainly been developed by Lazarsfeld (Lazarsfeld, 1950a, 1950b; Lazarsfeld & Henry, 1968) whereas log-linear models with latent variables have been mainly introduced by Goodman (1974a, 1974b), Habermann (1979), McCutcheon (1987), and Hagenaars (1990, 1993). Hagenaars incorporated more than one latent variable into the log-linear model with latent variables in his “modified LISREL approach”. Hagenaars (1990, 1993) based his approach on the theory of modified path models (Goodman, 1973). He showed how log-linear models with latent variables can be used to analyze directional relations between latent and manifest categorical variables (Hagenaars, 1990, 1993).

The two modeling strategies (LCA modeling and log-linear models with latent variables) can be seen as the categorical counterpart of metric or ordinal structural equation modeling (SEM). These models are based on extensions of the basic log-linear model (Goodman, 1974a, 1974b; Haberman, 1979; McCutcheon, 1987) and the LCA model (Lazarsfeld, 1950a, 1950b; Lazarsfeld & Henry, 1968) to log-linear models with latent variables. In fact, LCA models can be seen as a special variant of log-linear models with latent variables. The parameters of both models can be transformed into one another. Maximum Likelihood (ML) estimation procedures exist for both models (Clogg, 1981;

Goodman, 1974a, 1974b; Haberman, 1976, 1977, 1979; Hagenaars, 1993; Langeheine & Rost, 1988; McCutcheon, 1987). However, the log-linear parameterization allows for a more flexible modeling, because the (conditional) response probabilities of the LCA model are decomposed into effects due to underlying one-variable effects and possible interactions between variables. To analyze MTMM models it is, thus, advantageous to use the broader frame of log-linear modeling. Defining log-linear models with latent variables as latent rater agreement models also allows for an inspection of (conditional) response probabilities and proportions. In some cases (boundary values, see Section 4.1.2), only the (conditional) response probabilities can be interpreted. Additionally, the special parameter restrictions (e.g., quasi-independence restrictions) of latent rater agreement models can better be handled in the log-linear modeling framework. Therefore, I will define the rater agreement models in the log-linear modeling framework.

4.1.1 Formal Definition of the Log-Linear Model with Latent Variables

Table 4.1.1 depicts parts of a frequency table of a joint distribution of four three-categorical items measuring neuroticism (see Section 4.1.3, for more details). The total joint distribution consists of 81 different frequency patterns (3^4 cells in the joint distribution). The log-linear model with latent variables aims at representing these 81 response patterns in a parsimonious way (with a smaller number of parameters than possible frequency patterns). Therefore, the population is supposed to consist of several (homogeneous) sub-groups (classes of the latent variable) each showing the same relations to the items (the same log-linear parameters). Since the log-linear parameters can be transformed into conditional response probabilities, the expected frequency / proportion of every response pattern can be determined.

In contrast to the log-linear models presented in the introduction, the models presented here contain observed (manifest) as well as unobservable (latent) variables. The latent variables are supposed to influence the expected score on the manifest variables. In the basic model, which will be presented in this section, all manifest and latent variables are considered nominal variables, whereas extensions of this approach may also contain ordinal or metrical variables (e.g., Heinen, 1993; McCutcheon, 1987). The latent variables of LCA are, generally, called latent class variable, typological variable or latent trait variable. All these terms will be used for latent variables representing particular constructs.

Table 4.1.1

Four observed response patterns for self-report data measuring neuroticism (extracted from the complete table in Appendix A)

A vulnerable	*B* Sensitive	*C* Moody	*D* self-doubtful	Frequency	Relative frequency
…	…	…	…	…	…
1	1	1	1	8	.02
1	2	2	1	1	.00
3	3	2	2	15	.03
3	3	3	3	111	.23
…	…	…	…	…	…

Note. 1: non-neurotic response category; 2: middle response category; 3: neurotic response category.

In all modeling approaches, items measuring the same construct are statistically linked to a variable representing exactly this psychological construct (Bock, 1972; Langeheine & Rost, 1988; Lazarsfeld & Henry, 1968; McCutcheon, 1987; Steyer & Eid, 2001). The items depicted in Table 4.1.1 are supposed to measure different categories of neuroticism and should, thus, be linked to a latent variable representing different types of neurotic personalities (e.g., neurotic individuals, non-neurotic individuals, and individuals being in the "middle" of the two extremes). The categorical trait is supposed to cause an individual’s responses to the manifest indicators. Depending on her or his value on this categorical trait (her or his latent class membership), there will be differences in the expected frequencies of the different response patterns. These differences depend uniquely on the latent status of the individuals (see Figure 4.1)

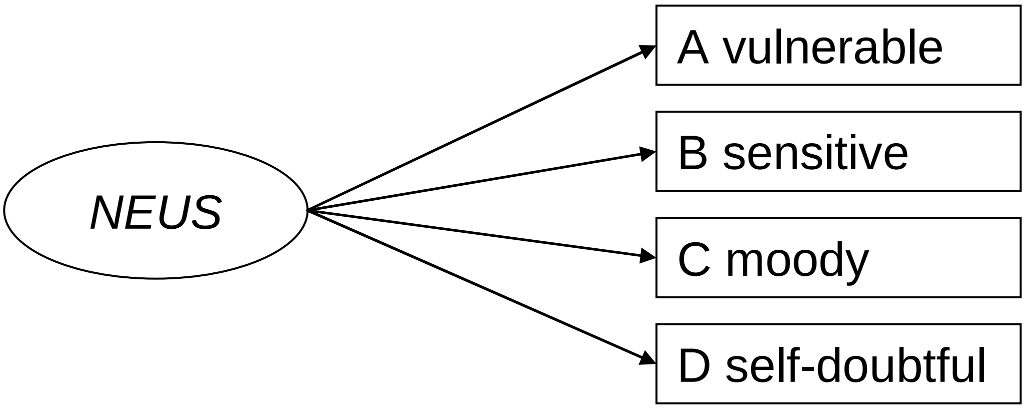

Figure 4.1. Basic log-linear model with one latent variable (*NEUS*) for neuroticism.

Definition 4.1.1 The log-linear model with one latent variable (see e.g., Hagenaars, 1990)

$$e_{\mathbf{a}.x} = \eta \mathrm{T}_{\mathbf{a}} \tau_x^X \qquad (4.1.1)$$

with $e_{\mathbf{a}.x}$ as expected frequency of the manifest response pattern $\mathbf{a}$ (e.g. [1 2 1 2]) given class membership *x*. η is the overall geometric mean of the complete table (manifest and latent variables).

$\mathrm{T}_{\mathbf{a}}$ represents the one-variable effects of the manifest variables and the two-variable effects linking the latent variable *X* to its indicators:

$$\mathrm{T}_{\mathbf{a}} = \prod_{M_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X} \, , \qquad (4.1.2)$$

with $\tau_{m_i}^{M_i}$ representing the one-variable effect for a category m of the *i*th item (out of the set of *I* items). *X* represents the latent variable and *x* the category of the latent variable. $\tau_{m_i.x}^{M_i.X}$ represents the two-variable log-linear effect of the latent category *x* on category *m* of item *i*. τ_x^X represents the latent one-variable effect (its latent distribution). Throughout this dissertation latent variables (categories) and their formal representations will be separated by a dot (“.”) from all other variables to discriminate them from the manifest variables (categories).

The log-linear model with one latent variable is defined in such a way that the manifest variables are independent from each other if the latent variable is controlled for. This is the condition of local stochastic independence. All associations between manifest variables are due to the their associations with the latent variable.

For the example presented in Figure 4.1, the log-linear model with latent variable is:

$$e_{abcd.ns} = \eta \tau_a^A \tau_b^B \tau_c^C \tau_d^D \tau_{ns}^{NEUS} \tau_{a.ns}^{A.NEUS} \tau_{b.ns}^{B.NEUS} \tau_{c.ns}^{C.NEUS} \tau_{d.ns}^{D.NEUS} , \qquad (4.1.3)$$

where *A* through *D* represent the manifest indicators of neuroticism, *a* through *d* the manifest categories of the corresponding indicators, *NEUS* is the latent variable representing neuroticism and *ns* are its categories. In the model described in equation 4.1.3: $\mathrm{T}_{\mathbf{a}} = \prod_{M_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.ns}^{M_i.NEUS} = \tau_a^A \tau_{a.ns}^{A.NEUS} \tau_b^B \tau_{b.ns}^{B.NEUS} \tau_c^C \tau_{c.ns}^{C.NEUS} \tau_d^D \tau_{d.ns}^{D.NEUS}$ with e.g., $\tau_{m_1}^{M_1} \tau_{m_1.ns}^{M_1.NEUS} = \tau_a^A \tau_{a.ns}^{A.NEUS}$.

4.1.1.1 The statistical meaning of the different effects in the log-linear model with one latent variable

The log-linear parameters of Definition 4.1.1 with unknown frequencies of the latent table (the cross-classification of observed and unobserved proportions) can be calculated as in the case of completely observed tables. Habermann (1979, p. 543) pointed out that “the same maximum likelihood equations apply as in the ordinary case, in which all frequencies are directly observed, except that the unexpected counts are replaced by their estimated conditional expected values given the observed marginal totals”. Thus, the estimated parameters have exactly the same meaning as in the ordinary model:

- η is the geometric mean of the unobserved complete frequency table (see e.g., Hagenaars, 1990). It is generally not of interest in models with latent variables.

- The latent one-variable parameter $\left(\tau_x^X\right)$ describes the univariate distribution of the latent variable. These parameters are identical to the odds comparing the probability (the proportion: π_x^X) of a particular category (x) with the geometric mean of all cells belonging to this variable (X):

$$\tau_x^X = \frac{\pi_x^X}{\sqrt[X]{\prod_{w=1}^{X} \pi_w^X}} . \tag{4.1.4}$$

with x and w indicating the categories of the latent variable X[9].

- The measurement model $\mathrm{T}_{\mathbf{a}} = \prod_{M_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$ depicts the relation of the manifest indicators to their underlying latent variable (the conditional response probability / conditional expected frequency). The model parameters of the measurement equation are based on the (unobserved) proportions (see Hagenaars, 1990):

$$\tau_{m_i}^{M_i} = \frac{\sqrt[X]{\prod_{x=1}^{X} \pi_{m_i.x}^{M_i.X}}}{\sqrt[IX]{\prod_{x=1}^{X} \prod_{n_j=1}^{I} \pi_{n_j.w}^{M_i.X}}} , \tag{4.1.5}$$

with j indicating the number of categories for item n. x and w indicating the categories of the latent variable X, and I indicating the number of categories for item n_j.

[9] X denotes the name of the latent variable as well as the number of categories. It only refers to the number of categories in connection with sum- or product signs $(\Sigma \text{ or } \Pi)$. The same is true for all other latent variables in this dissertation.

4.1.1.2 Conditional response probabilities in the log-linear model with latent variables

The log-linear models with latent variable can also be represented in two other parameterizations. All parameterizations can be transformed into each other. Equations 4.1.6 and 4.1.7 can be used to transform the log-linear parameters in proportions and conditional response probabilities (see e.g., Formann, 1992; Haberman, 1979; Heinen, 1996):

$$\pi_x^X = \frac{\tau_x^X}{\sum_{w=1}^{X} \tau_w^X} . \qquad (4.1.6)$$

with *w* indexing the different categories of *X*.

The conditional response probability to receive a particular response *m_i* on item *M_i* given that an individual belongs to latent category *x* can be determined:

$$\pi_{m_i.x}^{M_i.X} = \frac{\tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}}{\sum_{n_i=1}^{I} \tau_{n_i}^{M_i} \tau_{n_i.x}^{M_i.X}} , \qquad (4.1.7)$$

with n_i indexing the categories of M_i.

4.1.1.3 Effect-parameters of the log-linear model with latent variables

The second alternative parameterization is the effect-parameter parameterization. Effect-parameters can be used to examine the strength of the indicators' link to the latent variable in a way closely related to the inspection of the conditional response probabilities. One may conclude that an indicator is a good indicator of a latent category if it shows one large (or very low) effect-parameter. Effect-parameters represent odds and odds ratios. Computing the $\left(\Omega_{1/2bcd.x}^{\bar{A}BCD.X}\right)$, for example :

$$\Omega^{\bar{A}BCD.X}_{1/2bcd.x} = \frac{\eta\tau^{A}_{1}\tau^{B}_{b}\tau^{C}_{c}\tau^{D}_{d}\tau^{X}_{x}\tau^{A.X}_{1.x}\tau^{B.X}_{b.x}\tau^{C.X}_{c.x}\tau^{D.X}_{d.x}}{\eta\tau^{A}_{2}\tau^{B}_{b}\tau^{C}_{c}\tau^{D}_{d}\tau^{X}_{x}\tau^{A.X}_{2.x}\tau^{B.X}_{b.x}\tau^{C.X}_{c.x}\tau^{D.X}_{d.x}} = \frac{\tau^{A}_{1}\tau^{A.X}_{1.x}}{\tau^{A}_{2}\tau^{A.X}_{2.x}} = \Omega^{\bar{A}.X}_{1/2.x}, \qquad (4.1.8)$$

determines if it is more probable $\left(\Omega^{\bar{A}BCD.X}_{1/2bcd.x} > 1\right)$ or less probable $\left(\Omega^{\bar{A}BCD.X}_{1/2bcd.x} < 1\right)$ to receive a response in the 1st category of manifest item A given latent status *x* compared to the 2nd category given the same latent status *x*. The latent score is fixed because one is interested in the ratio within exactly this category of the latent variable. Parameters which do not contain the superscript of the manifest variable of interest (e.g., *A* - “vulnerable”) can be cancelled because their categories are held constant. The complex multi-way (3x3x3x3x3) contingency table can thus be represented in several subtables which only consist of the latent variable and one manifest variable. It is possible to collapse across all other manifest variables because all manifest variables are independent from each other given the latent variable (see Bishop, 1971; Appendix B). The ratio $\Omega^{\bar{A}BCD.X}_{1/2bcd.x}$, with the simplified notation of $\Omega^{\bar{A}.X}_{1/2.x}$, consists of two components, representing the main effect of the manifest variable and the interaction term:

$$\Omega^{\bar{A}.X}_{1/2.x} = \frac{\tau^{A}_{1}}{\tau^{A}_{2}} \times \frac{\tau^{A.x}_{1.x}}{\tau^{A.X}_{2.x}} = \gamma^{\bar{A}}_{1/2}\gamma^{\bar{A}.X}_{1/2.x}, \qquad (4.1.9)$$

with $\gamma^{\bar{A}}_{1/2} = \frac{\tau^{A}_{1}}{\tau^{A}_{2}}$ representing the general effect to be rather in the first than in the 2nd class and $\gamma^{\bar{A}.X}_{1/2.x} = \frac{\tau^{A.X}_{1.x}}{\tau^{A.X}_{2.x}}$ represents the change in the general effect $\left(\gamma^{\bar{A}}_{1/2}\right)$ as a function of the latent category. One may also calculate the odds to choose the 1st rather than the 2nd or $(\vee)$ 3rd category $(2 \vee 3)$:

$$\Omega^{\bar{A}BCD.X}_{1/(2\vee3)bcd.x} = \frac{\tau^{A}_{1}\tau^{A.X}_{1.x}}{\sum\limits_{a=2}^{3}\tau^{A}_{a}\tau^{A.X}_{a.x}} = \Omega^{\bar{A}.X}_{1/(2\vee3).x}\,. \qquad (4.1.10)$$

4.1.1.4 Implications of the log-linear model with one latent variable

The standard log-linear model with one latent variable (LCA model) serves as the smallest sub-model of the latent rater agreement models and the Multitrait-Multirater models. It is the measurement model for the latent trait variables. The model and its three parameterizations serve to identify the reliability of the indicators and the meaning of the latent variable. Hagenaars (1993) pointed to a parallel between models for continuous variables and the models presented here. The direction and the strength of the link between the latent variable and its indicators mainly serve to determine the meaning of the latent variable in models with unordered categorical latent. Analyzing the meaning of the latent variable is nothing else than examining its validity and / or the validity of the measures (e.g., Messick, 1989). The validity of a measure has its upper bound in the reliability. There are three ways to inspect the reliability of an indicator:

1. High two-variable log-linear parameters indicate if a manifest category is linked to a latent category. However, these parameters cannot be interpreted on their own but have to be compared across the latent categories. This comparison is more easily done relying on the effect-parameters (see below).
2. The reliability can also be determined by the inspection of the conditional response probabilities. If all conditional response probabilities of different indicators point to one specific manifest category as a function of the latent variable, there is an indication of reliability. That is, all manifest categories of all indicators supposed to measure a neurotic personality type, for example, have positive effects between the latent and manifest categories representing this type (matching categories).

   Dillon and Mulani (1984) as well as Langeheine (1988) present how the conditional response probabilities can be used to determine the classification errors of different raters rating one target on one manifest variable[10]. The classification errors are the weighted (by class sizes) sum of classification errors (see Langeheine, 1988). The inverse of the classification errors quantifies the reliability of an indicator.

   However, their approach does not apply to all cases of latent rater agreement models. It requires that all items represent the same content and that the categories of these items correspond to one and only one category. This can only be adapted to the analysis of multiple indicators if all categories of the indicators represent the same

[10] The different raters are treated as indicators are treated in the approaches of Dillon and Mulani (1984) or Langeheine (1988).

contents. This does not necessarily have to be the case for multiple indicators of one construct. Consider the items “vulnerable” and “moody” as indicators of neuroticism. It may turn out that individuals being moderately neurotic are highly “vulnerable” but may be more or less “moody” without preferring a special response category for this item. In the sense of Dillon and Mulani, item “vulnerable” is highly reliable whereas item “moody” shows low reliability. However, item “moody” may be needed to differentiate between moderately and highly neurotic individuals. Moderately as well as highly neurotic individuals are highly “vulnerable” but only highly neurotic individuals are also highly “moody”.

3. Determining the effect-parameter for every manifest category reveals, if there is one special manifest category which can be seen as a marker for the latent category. Very high effect-parameters indicate that it is much more probable to choose this category than one of the other categories.

   If all two-variable effect-parameters point to the same direction for every latent category, respectively, one may additionally examine their absolute values. If the manifest one-variable effect-parameters as well as the two-variable effect-parameters show identical values for two indicators, these indicators can be considered homogeneous. Like in models for homogeneous raters (Schuster, 2002; Zwick, 1988) homogeneity is not only reflected in equivalent two-variable effects, but also in equivalent manifest one-variable effects. In this case, the model predicts the same manifest distribution for the indicators. This allows for a test if all indicators share the same categories representing the latent traits.

   If the categories differ with respect to their effect parameters their categories represent different latent statuses. In the case of ordered latent categories, for example, one category (e.g., sometimes) may be the typical response tendency for a high latent status on a particular construct (say depression) if the item describes a rare behavior (e.g., “do you wish to be dead?”) but also a typical response category of an easy item for a low status on the same construct (e.g., “do you feel helpless?”).

4. The mean assignment probabilities could also be used to determine the reliability of the latent categorization based on the items. This coefficient indicates the mean probability to be assigned to the class an individual most probably belongs to. That is, if an individual has the relatively highest probability to belong to class *x*, she or he will

be assigned to this class. The mean assignment probability is the mean of all assignment probabilities of all individuals who are assigned to this class[11].

4.1.2 Estimation Process and Boundary Values

The estimation of log-linear models with latent variables cannot be done using analytic strategies. Instead Maximum Likelihood (ML) estimation procedures have to be used. The most common procedures use either the Expectation-Maximization (EM) algorithm, particular variants of the Newton/Raphson procedure, or a combination of these approaches (Galindo-Garre & Vermunt, 2004, 2005, 2006; Goodman, 1974a, 1974b; Haberman, 1979, 1988; Hagenaars, 1990). Iterative proportional fitting (IPF) procedures can be used to find the expected frequencies e of hierarchical log-linear models without latent variables (Fienberg, 1980; Hagenaars, 1990). In IPF the initial estimates E are iteratively adapted, so that they finally fit the observed marginal frequencies *f*. The algorithm, thus, aims to reproduce the observed marginal distributions. Models with latent variables are estimated in a similar way. However, as the latent variables cannot be observed, the EM algorithm has to be used in order to reproduce the observed frequency table.

One problem with this estimation method is, that in some cases parameter estimates may occur that are on the edges of the parameter space (boundary solutions). These boundaries correspond to probabilities of $\pi = 0$ or $\pi = 1$ and to $\tau = 0$ or to undefined τ-parameters as values of log-linear parameters. Boundary values may be due to the following reasons:

1. *Empirical non-identification.* Large probability tables with relatively small samples (Winship & Mare, 1989). This situation is also called sparse table problem. This problem may principally be solved increasing the sample size.
2. *Intrinsic non-identification.* This case to produce boundary solutions can occur in cases where many solutions exist for the set of model equations. Repeated analyses of the same model will yield different results (see e.g., Formann, 1992; Galindo-Garre & Vermunt, 2004, 2005, 2006; Goodman, 1974b; McCutcheon, 1987; Winship & Mare, 1989).

[11] These coefficients are not provided by software package LEM (Vermunt, 1997a) which will be used for the empirical analyses.

3. *Structural zeros and true parameters*. In some applications it is meaningful to find conditional response probabilities of 1 or 0 (Galindo-Garre & Vermunt, 2004, 2005, 2006). There should be no male taking the birth control pill and thus the response probability for males on this item will be a structural zero (a cell that cannot be observed). However, it is the true parameter because males do not take this pill.

In all these cases, log-linear parameters cannot be interpreted because they are not identified. Additionally, if the model design comprises more parameters than observed response patterns minus 1, the model is underidentified, which implies that there is an infinite number of “best” solutions of the estimation process (see e.g., Formann, 1992; Galindo-Garre & Vermunt, 2004, 2005, 2006; Goodman, 1974b; McCutcheon, 1987; Winship & Mare, 1989).

Boundary values lead to numerical problems in the computation of the parameters' variance-covariance matrix and to meaningless confidence intervals and significance tests (see Galindo-Garre & Vermunt, 2004, 2005, 2006). If there are boundary values, the inverse of the information matrix cannot be determined and thus no standard errors can be calculated. The standard errors of the non-boundary parameters can be calculated taking the generalized inverse of the information matrix. These standard errors are only valid, if the boundary parameters are considered true (a priori) model parameters (see Galindo-Garre & Vermunt, 2004, 2005, 2006). Model probabilities still can be interpreted if boundary values have been found, yet, log-linear and effect-parameters are not interpretable (dividing by zero is not defined).

There have been different attempts to solve the different problems of boundary solutions. De Menezes (1999) proposed to use the parametric bootstrap to overcome the problems of meaningless standard errors. Her results show that the bootstrap procedure yields accurate estimates for the conditional response probabilities; yet, she could not solve the problem that boundary solutions may occur during the bootstrap procedure yielding invalid bootstrap results for the effect-parameters. Maris (1999) used prior information on the model parameters and thus a Bayesian estimation method called posterior mode or maximum a posteriori estimation. Unfortunately, this method is not available in the software package LEM (Vermunt, 1997a), which will be used in the empirical applications. It is available in Latent GOLD (Vermunt & Magidson, 2000, 2005), however, Latent Gold does not allow for more than one latent class variable. Therefore, these newly developed estimation methods will not be discussed further.

Although the interpretative problems with respect to effect-parameters have been well known for a long time now, there is no common sense how to deal with them in empirical applications. The majority of research groups seem to argue that the model probabilities still can be interpreted. There is agreement that the effect-parameters should not be interpreted. Yet, there are different points of views concerning the consequences for the degrees of freedom (*df*). Some authors add the number of parameters on the boundary to the number of the degrees of freedom (e.g., McCutcheon, 1987). Others state, that they see no good reason to do so (Magidson & Vermunt, 2001). In the remainder, I will consider all parameters (including parameters on the edge of their parameter space) as model parameters and account for them in reporting the degrees of freedom (as is done in LEM; Vermunt, 1997a).

4.1.3 Application of the Log-Linear Model with One Latent Variable

In order to illustrate the models to be developed in this dissertation all models will be applied to empirical data. I will use the same data-set with changing constellations of raters. Therefore, the complete data set is described now. For every application, I will explicitly list the raters and variables that will be analyzed.

4.1.3.1 Data description

The data used in this dissertation originate from a large study conducted by Eid, Lischetzke, Nussbeck, and Geiser (2004) at the University of Trier (Germany) in 2001 and 2002. Out of the about 15000 students a random sample of 3000 students was sent a mail inviting them to come to the laboratory bringing two peers along. The student who received the mail and who came to the laboratory was asked to fill in a self-report questionnaire (target person: *S*) and the two peers were asked to fill in the same questionnaire but in the peer-report version (peer *A* and peer *B*). As originally intended, data from 500 students could be finally collected extending the study to the University of Applied Sciences (FH) at Trier. The study yields a data set of 500 triples (1500 individuals in total).

All three members of the triple were asked to fill in the questionnaires. The participants were separated to prevent them from sharing information. Filling in the complete questionnaire took about 30-45 minutes and every participant received a compensation of 20 German Marks (DM). Each participant was allowed to participate only once (irrespective if as target person or as peer *A* or *B*). Although all participants were informed about this restriction and signed a receipt confirmation for the compensation (including their address) 17 triples could be identified with individuals who participated twice. That triple was eliminated where the person participated for the second time. Another triple was eliminated containing a participant who was only 13 years old. For the analyses presented in this dissertation only complete data sets will be used. Therefore, four more triples had to be excluded because their data sets yielded missing data. The final data set thus contains data from 478 triples; that is, 1434 participants.

4.1.3.2 Sample description

About two third of all participants are female students (63.7% of the target persons, 62.9% of peers A, 62.9% of peers B). Thus, women are slightly overrepresented in the sample with respect to the proportion of enrolled female students at Trier Universities (about 55% of the students are female; for a more detailed discussion see Nussbeck, 2002). The sample consists mainly of students studying one of the following five subjects: Psychology, Economics, Law, Architecture, and Geography / Geology (see Table 4.1.2)

Table 4.1.2

Sample description with respect to the most frequently studied subjects

Subject	Proportion of all enrolled students at Trier Universities[a]	Proportion of female students in this program[a]	Proportion of students enrolled in this program in the sample	Proportion of female students enrolled in this program in the sample
Psychology	9,9%	69,5%	17,1%	76,1%
Economics	14,7%	42,2%	15,6%	58,9%
Law	17,3%	54,9%	15,0%	69,1%
Architecture (FH)[b]			7,6%	56,4%
Geography/Geology	12,6%	54,0%	7,1%	66,7%

Note. [a] Data stem from fall 1999/2000; [b] Unfortunately, no statistics were available for the University of Applied Sciences. The percentages do not sum up to 100% because not all subjects are listed.

The mean age of all participants (target persons and peers) is 23.4 years. The youngest participants were 17 years old for target persons and peers *A*, the youngest peer *B* was 18 years old, the oldest participants were 49 years old for target persons and peers *A*, and 52 years old for peers *B*. About 66% of all participants are between 19 and 27 years old (corresponding to the expectations about a student sample; see Table 4.1.3). The sample (in all three groups) was highly qualified since more than 98% of all participants had at least "Fachhochsschulreife" (high school degree which permits attending German Universities of Applied Sciences: FH). More than 94% of the total sample was enrolled at the University or at the University of Applied Sciences.

Table 4.1.3
Sample description with respect to age

	Target Person	Peer A	Peer B
Mean	23,4 years	23,4 years	23,4 years
1. Percentile 0-25%	17-21 years	17-21 years	17-21 years
2. Percentile 25-50%	21-23 years	21-23 years	21-23 years
3. Percentile 50-75%	23-25 years	23-25 years	23-25 years
4. Percentile 75-100%	25-49 years	25-49 years	25-52 years
Youngest	17 years	17 years	18 years
Oldest	49 years	49 years	52 years

Table 4.1.4
Time the target person S knows peers A and B

Percentile	Time *S* knows *A*	Time *S* knows *B*
1. Percentile 0-25%	up to 5 month	up to 5 month
2. Percentile 25-50%	5-20 month	5-18 month
3. Percentile 50-75%	20-42 month	18-38 month
4. Percentile 75-100%	42-311 month	38-294 month

The target person and peers *A* and *B* know each other fairly well. On a 10-point scale (10 indicating best knowledge / highest familiarity: “We have absolutely no secrets”) the target persons have a mean value of 6.51 for the familiarity with *A* and 6.44 for the familiarity with *B*. Peers *A* indicate a mean value of 6.64 and *B* of 6.59 for the familiarity with the target person. The intraclass correlations (*ICC*) for these variables are *ICC* = .82 (target person and A) and *ICC* = .78 (target person and B). Target persons and peers, thus, rate their familiarity on a relatively high level and very similar to each other. The time each dyad knows each other is depicted in Table 4.1.4. On average, the target person and the peers have known each other for three years (*ICC* = .96 for the target person and peer *A* and *ICC* = .99 for the target person and peer *B*). Target persons and peers thus agree (almost) perfectly about the time they know each other. 75% of the participants have known each other for at least half a year, 50% for at least more than one and a half years (see Table 4.1.4). With respect to familiarity time the dyads know each other, they do virtually not differ from each other and, therefore, the two peer raters should be able to judge the target's traits being considered interchangeable.

4.1.3.3 Variables

In this dissertation, two sub-scales of a German Big-Five scale (Ostendorf, 1990) measuring neuroticism and conscientiousness are used to illustrate all newly developed models. Neuroticism and conscientiousness were selected because prior research result showed that facets of conscientiousness (being dependable) enhanced rater agreement and facets of neuroticism (being moody) deteriorated rater agreement (see Colvin, 1993b). The scales in the self-report version can be found in Appendix A. The response format in its current form is an ordered response format ranging from "not at all" to "very much so" across five categories (see Appendix A). Therefore, the data could principally be analyzed using dimensional models for ordinal response formats [i.e., models of Item Response Theory (IRT); Andrich, 1978; Jansen & Roskam, 1986; Roskam, 1995; Roskam & Jansen, 1989; Samejima, 1969].

Since the aim is to develop MTMR models for categorical and non-ordered categorical response variables the range of the scale was reduced to three categories in order to reduce the complexity of the model. Analyzing log-linear models with variables consisting of 5 categories will result in 5^I possible manifest response patterns, where I indicates the number of items. In order to have models that do not suffer from empirical non-identification by default (due to the large number of possible patterns) the extreme categories were collapsed: The first and second categories (the lowest categories) have been collapsed, the middle category has been kept, the fourth and fifth categories (highest categories) have been collapsed. Still, the three remaining categories are ordered. The frequency distributions of the 8 items for the self-report (target person) can be found in Table 4.1.5. The frequency distributions of the two peer reports *A* and *B* are quite similar and can be found in Appendix A.

Table 4.1.5

Frequency distribution of the analyzed Big-Five Items (self-report)

		Categories			
German item	English item	little (1)	middle (2)	highly (3)	total
		neuroticism			
verletzbar	vulnerable	43	75	360	478
empfindlich	sensitive	63	77	338	478
launenhaft	moody	179	130	169	478
selbstzweiflerisch	self-doubtful	121	88	269	478
		conscientiousness			
arbeitsam	industrious	93	165	220	478
fleißig	diligent	116	159	203	478
pflichtbewußt	dutiful	29	93	356	478
strebsam	ambitious	122	150	206	478

Note. Categories 1 and 2 as well as 3 and 4 of the original scale have been collapsed.

The two peer raters have been randomly assigned to be peer *A* or *B*, they can be conceived interchangeable. This assumption seems to be tenable because the two peers do not differ with respect to the distribution of their variables presented in Tables 4.1.3 and 4.1.4 (see Schuster, 2002; Schuster & Smith, 2002, 2006; Zwick, 1988). The two peer reports differ structurally from the self-report.

4.1.3.4 Application of the log-linear model with one latent variable

To illustrate the log-linear model with one latent variable I will present the results of this model in detail. In this section the four self-report items ("vulnerable, sensitive, moody, and self-doubtful") will be analyzed. All models in this dissertation are estimated using the software package LEM (Vermunt, 1997a). The corresponding input files can be found in Appendix F.

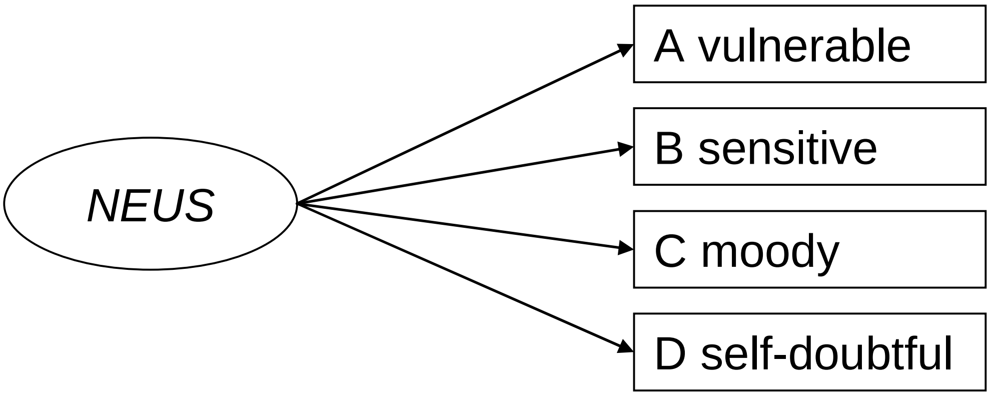

Figure 4.1 (repeated). Basic log-linear model with one latent variable (*NEUS*) for neuroticism.

The χ^2-criterion and the information criteria AIC (Akaike, 1974, 1987; Bozdogan, 1987) and BIC (Schwartz, 1978) will be used to evaluate the goodness-of-fit of the different models. Additionally, I will run bootstrap analyses with *N* = 200 bootstrap samples to check for the overall goodness-of-fit (relying on the simulated Pearson-χ^2-values) because the models are very likely to produce empirical χ^2- values that do not approximate the theoretical distribution (sparse table problems, see Habermann, 1988; Hagenaars, 1990; Winship & Mare, 1999).

Figure 4.1 depicts the log-linear model with one latent variable for the empirical application. The latent variable *NEUS* for neuroticism in the self-report underlies the manifest response behavior. Latent variables are presented with ovals, manifest variables with boxes, arrows indicate dependencies between variables. In this approach, no error components are depicted, because the dependent variables do not correspond to the responses themselves but expected frequencies of each category for every indicator. The response depends uniquely on the latent variable *NEUS*. This is the assumption of local stochastic independence.

Table 4.1.6

Goodness-of-fit coefficients of the log-linear model with 1, 2, and 3 latent categories

NS	χ^2	$p\left(\chi^2\right)$	L^2	$p(L^2)$	*df*	AIC[1]	BIC[1]	p_{boot}	n_{bounds}
1	562.25	.00	356.09	.00	72	212.09	–88.12	—	—
2	79.62	.08	86.45	.03	63	–39.55	–302.23	.09	1
3[2]	59.61	.28	65.27	.14	54	–42.73	–267.89	.38	7

Note. NS: number of latent categories; χ^2: Pearson χ^2-value; L^2: Likelihood–Ratio χ^2-value; [1]AIC and BIC are based on the L–squared χ^2–value; [2] The estimation of the three–class solution yielded one fitted zero marginal in LEM; p_{boot}: bootstrapped probability of χ^2; n_{bounds}: number of boundary values.

Table 4.1.6 shows the goodness-of-fit indices for the three different models. The one-class solution does not fit to the data. The two-class solution fits to the data according to the χ^2-value, does not fit with respect to the L^2 value, and fits with respect to the bootstrapped χ^2-value. The three class solution generally fits to the data according to these three criteria. According to the AIC, the three-class solution should be preferred. According to the BIC the two-class solution should be preferred.

The three-class solution will be presented because the latent rater agreement models that will be defined in Section 5 require at least three latent categories to differentiate from one another. I will exemplarily report all three parameterizations to illustrate their meanings. In the remainder, I will mostly rely on the conditional response probabilities to illustrate the relation between the manifest and the latent variables since the conditional probabilities can be interpreted even in the case of boundary solutions. However, in some cases I will also refer to the other parameterizations.

Table 4.1.7

Multiplicative log-linear parameters of the log-linear model with three latent categories representing neuroticism (self-report)

Overall effect				
$\hat{\eta}$	$1.15\ 10^{-28}$			
One variable effect of the categorical trait				
	ns = 1	*ns* = 2	*ns* = 3	
$\hat{\tau}_{ns}^{NEUS}$	$1.59\ 10^{11}$	$3.45\ 10^{27}$	$1.61\ 10^{-37}$	
One–variable effects of the manifest variables				
	r = 1	*r* = 2	*r* = 3	Variable names
$\hat{\tau}_{a}^{A}$	$2.50\ 10^{-41}$	$4.40\ 10^{19}$	$9.08\ 10^{20}$	“vulnerable”
$\hat{\tau}_{b}^{B}$	$1.22\ 10^{8}$	$1.36\ 10^{5}$	$2.96\ 10^{-11}$	“sensitive”
$\hat{\tau}_{c}^{C}$	0.73	0.37	3.68	“moody”
$\hat{\tau}_{d}^{D}$	0.47	0.91	2.36	“self–doubtful”
Two–variable effects of the latent variable and its indicators				
	ns = 1	*ns* = 2	*ns* = 3	
$\hat{\tau}_{1.ns}^{A.NEUS}$	$3.52\ 10^{40}$	$2.86\ 10^{-20}$	$9.92\ 10^{-22}$*	“vulnerable”
$\hat{\tau}_{2.ns}^{A.NEUS}$	$9.83\ 10^{39}$	$1.58\ 10^{-20}$	$6.44\ 10^{-21}$*	
$\hat{\tau}_{3.ns}^{A.NEUS}$	$2.89\ 10^{-81}$	$2.21\ 10^{39}$	$1.56\ 10^{41}$	
$\hat{\tau}_{1ns}^{B.NEUS}$	$6.64\ 10^{8}$	$5.95\ 10^{11}$	$5.09\ 10^{-24}$*	“sensitive”
$\hat{\tau}_{2.ns}^{B.NEUS}$	$1.50\ 10^{-9}$	$4.83\ 10^{-6}$	$2.76\ 10^{11}$*	
$\hat{\tau}_{3.ns}^{B.NEUS}$	0*	$3.48\ 10^{-7}$	$7.11\ 10^{11}$	
$\hat{\tau}_{1.ns}^{C.NEUS}$	3.63	2.21	0.12	“moody”
$\hat{\tau}_{2.ns}^{C.NEUS}$	1.59	3.30	0.19*	
$\hat{\tau}_{3.ns}^{C.NEUS}$	0.17	0.13	42.12	
$\hat{\tau}_{1.ns}^{D.NEUS}$	3.37	1.11	0.27	“self–doubtful”
$\hat{\tau}_{2.ns}^{D.NEUS}$	1.83	0.59	0.93	
$\hat{\tau}_{3.ns}^{D.NEUS}$	0.16	1.52	4.04	

Note. * boundary values. 1 fitted margin is zero. *ns*: latent category; *r*: manifest category.

*Log-linear parameter*s. The estimates of the population parameters (marked with a hat) depicted in Table 4.1.7 should only heuristically be interpreted because they are afflicted by boundary values. The first row presents overall geometric mean $\left(\hat{\eta}\right)$. This parameter is a mere reflection of the sample size (Hagenaars, 1990). The log-linear parameters for the latent variable (latent one-variable effect: $\hat{\tau}_{ns}^{NEUS}$) show that the middle category is strongly preferred by the raters followed by the 1st category, the 3rd latent category is not preferred according to this parameters. The log-linear parameters of the manifest distribution (manifest one-variable parameters, e.g.: $\hat{\tau}_{a}^{A}$) depict the unconditional manifest distribution. It can be seen that the 2nd and 3rd category are strongly preferred for the item "vulnerable" (*A*). The first two categories of item "sensitive" (*B*) are more frequently expected than based on the geometric mean. The manifest log-linear parameters for items "moody" (*C*) and "self-doubtful" (*D*) show that the 3rd category is overrepresented for these items.

The two-variable log-linear parameters $\left(\text{e.g., } \hat{\tau}_{1.ns}^{A.NEUS}\right)$ show that the link between the 1st latent class (ns = 1) and the 1st manifest response category is always strongest because the two-variable log-linear parameter is highest for this connection. However, for three items (*A*, *C* and *D*) also the 2nd manifest response category is strongly related to the 1st latent category. Principally, the parameter estimates decline with an increase in the index of the manifest category. The two-variable parameters linking the 2nd latent category to the manifest response categories reveal that this category is strongly linked to the 3rd manifest response category of item *A*, to the 1st manifest response category of item *B*, to the 2nd manifest response category of item *C*, and also linked to the 1st and 3rd manifest response category of item *D*. The two-variable parameters linking the 3rd latent category to the items are always highest for the 3rd manifest response category. The boundary values strongly influence the parameter estimates as shown by the very high values.

Effect parameters. The effect-parameters (see Table 4.1.8) give a more comprehensive view on the relation between the latent and the manifest variables because the manifest one-variable effects are considered in addition to the two-variable effects. The values are only depicted for items *C* and *D* because these parameters suffer from the boundary values to a smaller degree than the parameters for items *A* and *B*.

Table 4.1.8

Effect-parameters (category against the two others) for the two indicators “moody” and “self-doubtful” in the log-linear model with three latent categories representing neuroticism

One variable effect of the manifest variables

	$r = 1$	$r = 2$	$r = 3$	Variable names
$\hat{\gamma}_c^{\bar{C}}$	0.53	0.14	13.54	“moody”
$\hat{\gamma}_d^{\bar{D}}$	0.22	0.83	5.57	“self-doubtful”

Two-variable effects of the latent variable and its indicators

	$ns = 1$	$ns = 2$	$ns = 3$	
$\hat{\gamma}_{1.ns}^{\bar{C}.NEUS}$	13.18	4.88	0.01	
$\hat{\gamma}_{2.ns}^{\bar{C}.NEUS}$	2.53	10.89	0.04	“moody”
$\hat{\gamma}_{3.ns}^{\bar{C}.NEUS}$	0.03	0.02	1774.09*	
$\hat{\gamma}_{1.ns}^{\bar{D}.NEUS}$	11.36	1.23	0.07	
$\hat{\gamma}_{2.ns}^{\bar{D}.NEUS}$	3.35	0.35	0.86	“self-doubtful”
$\hat{\gamma}_{3.ns}^{\bar{D}.NEUS}$	0.03	2.31	16.32	

Note. * boundary value. *r*: manifest response category; *ns*: latent category. For sake of simplicity, the effect parameters are simplified to $\hat{\gamma}_{c.ns}^{\bar{C}.NEUS}$, for example, leaving out the indices for the two other categories.

Table 4.1.8 shows the effect-parameters to choose one particular category against the two other categories. As already described for the log-linear parameters, it is much more probable to choose the 1st manifest category ($\hat{\gamma}_{1/(2\vee3).1}^{\bar{C}.NEUS} = 13.18$ and $\hat{\gamma}_{1/(2\vee3).1}^{\bar{D}.NEUS} = 11.36$) if the target belongs to the 1st latent class. The effect-parameters for the 2nd latent class indicate that individuals belonging to this class have the highest tendency to choose the 2nd manifest response category for item *C* and the 3rd category for item *D*. However, the 1st manifest response category is also more often chosen than predicted by the one-variable effect-parameters $\left(\hat{\gamma}_{1/(2\vee3).2}^{\bar{C}.NEUS} > 1 \text{ and } \hat{\gamma}_{1/(2\vee3).2}^{\bar{D}.NEUS} > 1\right)$. Individuals belonging to the 3rd latent class most probably endorse the 3rd manifest category ($\hat{\gamma}_{3/(1\vee2).3}^{\bar{C}.NEUS} = 1774.09$ and $\hat{\gamma}_{3/(1\vee2).3}^{\bar{D}.NEUS} = 16.32$).

Conditional response probabilities. The analysis of the conditional response probabilities in Table 4.1.9 shows that the three latent categories can be interpreted as three latent personality types. A non-neurotic type ("non-neurotic" class: 1), a class (2) preferring neurotic response tendencies with respect to items *A* and *B* (“vulnerable” and “sensitive”) and showing no strong response tendencies for items *C* and *D* (“moody” and “self-doubtful”) - I will call this class "sensitive but stable class" to have a short description -, and a “neurotic” type (“neurotic” class: 3) choosing the third category with very high probabilities for all items.

Table 4.1.9
Conditional response probabilities in the log-linear model with three latent categories representing neuroticism

		latent categories[1]		
variable	manifest categories	1 $\left(\hat{\pi}_1^{NEUS} = .24\right)$	2 $\left(\hat{\pi}_2^{NEUS} = .56\right)$	3 $\left(\hat{\pi}_3^{NEUS} = .20\right)$
A (vulnerable)	1	.29	.04	.00*
	2	.41	.10	.00*
	3	.30	.86	.99
B (sensitive)	1	.50	.02	.00*
	2	.50	.07	.00*
	3	.00*	.91	.99
C (moody)	1	.68	.38	.00
	2	.21	.40	.00*
	3	.12	.23	.99
D (self-doubtful)	1	.49	.24	.01
	2	.31	.15	.13
	3	.20	.61	.87

Note. * boundary values. [1] The values in parentheses represent the latent class sizes. One fitted margin is zero.

The first class consists of 24% of all participants. The probability of choosing the 3rd response category are rather low (*p* = .30, .00, .12 and .20, respectively) and the probabilities to choose the second response category are not very pronounced (*p* = .41, .50, .21, and .31, respectively). Slightly more pronounced are the conditional response probabilities to choose the 1st manifest category for individuals belonging to the 1st class (*p* = .29, .50, .68, .49). Keeping in mind, that the categories of the variables are ordered from low neurotic to high neurotic response categories, the first class shows low neurotic response tendencies.

Individuals of the second class tend to choose the first response category much less often (probabilities of .04, .02, .38 , and .24), they also do not choose the second response category very often (probabilities of .10, .07, .40 , and .15), but rather tend to choose the third response category (probabilities of .86, .91, .23 , and .61); the typical response pattern for this class is to approve items *A* and *B* (choose the 3rd category) to choose any category for item *C* slightly preferring the categories 1 and 2 $\left(\hat{\pi}_{1.2}^{C.NEUS}=.38;\hat{\pi}_{2.2}^{C.NEUS}=.40;\hat{\pi}_{3.2}^{C.NEUS}=.23\right)$ and to most probably choose the highest response category for item *D* (self-doubtful) but to also choose another response category in 40% of the times. This type could best be described as a "sensitive but (emotionally) stable" [vulnerable and sensitive but not very moody or self-doubtful] personality type.

Members of the third latent class choose the third response category almost with certainty for the first three items ("vulnerable", "sensitive", and "moody", all *p* = .99) and strongly prefer the third category of item *D* ("doubtful" *p* = .87). This class shows clear neurotic response tendencies. The differences in terms of the typical response behaviour between the second and the third class of individuals mainly consist in the expected responses for "moody" and "self-doubtful". Individuals belonging to the 3rd class will mainly choose the third response category (86% of the time), whereas individuals belonging to the 2nd class will also provide responses in the 1st or 2nd response category for at least one item in 85% of the time. The empirical application shows that the interpretation of the conditional response probabilities is much clearer (with respect to possible differences between the effects of latent categories on the manifest indicators) than the interpretation of the effect-parameters.

Reliability. The log-linear parameters and the effect-parameters can only heuristically be analyzed because their parameters are afflicted by boundary estimates. However, the log-linear and effect parameters indicate that there is an ordered structure for the latent variable. The conditional response probabilities indicate if one latent category is strongly related to a specific response tendency for a manifest variable. This is generally the case for the 3rd latent category, but this is not the case for the other latent categories, except for items *A* and *B* for the 2nd latent category. Since identical named categories of the manifest variables do not correspond to identical latent categories the approach of Dillon and Mulani (1984) to inspect reliability relying on the conditional response probabilities may not be used. Yet, the mean assignment probabilities (determined by a run of this model in Mplus, Muthén & Muthén, 2007)[12] for the three-class solution are all above .78 (.79; .89; .90) indicating a reliable classification of individuals into the three classes.

4.2 Extension to More than One Latent Variable – Correlated Traits

As described in Section 1, the analyses of the convergent and discriminant validity can be done using the (CFA-) Correlated-Trait (CT) model. In this model, two or more traits are measured by multiple indicators administered to multiple raters. There is one latent variable for each Trait-Method-Unit (TMU). A TMU consists of all manifest ratings of one rater for one specific trait.

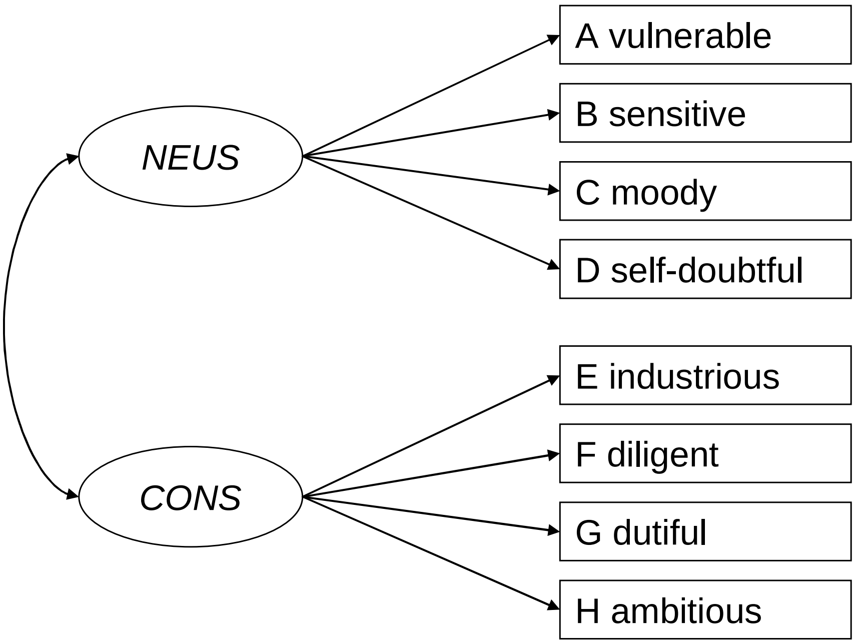

Figure 4.2. The loglinear-model with two latent variables for neuroticism and conscientiousness. *NEUS*: Neuroticism; *CONS*: Conscientiousness (self-report data)

[12] Mplus does not allow for estimations of more complex models. Therefore, I will only rely on the empirical results provided by LEM in the remainder.

In the framework of log-linear models with latent variables, additional latent variables can be easily incorporated (see e.g., Hagenaars, 1993). Figure 4.2 depicts a log-linear model with two latent variables. In this model, the two latent variables are measured by four manifest variables each. The double-headed arrow indicates that the two latent variables may be associated.

Definition 4.2.1

The log-linear model with two latent variables.

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \tau_{x.y}^{X.Y} \qquad (4.2.1)$$

is a log-linear model with two latent variables. $e_{\mathbf{ab}.ns.cs}$ is the expected frequency of a specific cell in the latent joint cross-classification of the manifest response patterns $\mathbf{ab}$ (consisting of the two trait-specific patterns $\mathbf{a}$ and $\mathbf{b}$) with the two latent variables *X* and *Y*. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the product of the log-linear parameters linking the latent variable *X* to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the product of the log-linear parameters linking the latent variable *Y* to its indicators and the manifest one-variable effects.

τ_x^X and τ_y^Y represent the latent one-variable effects. $\tau_{x.y}^{X.Y}$ represents the latent two-variable effects.

4.2.1.1 The statistical meaning of the different effects in the CT model

The log-linear parameters of Equation 4.2.1 with unknown frequencies of the latent table can be calculated as in the case of completely observed tables. Calculating the log-linear parameters of the sub-models for each trait can be done using the collapsed latent frequency table for each TMU. Since there is no interaction between the items being indicators of one trait and the items being indicators of the other trait, the collapsibility theorem holds (Bishop, 1971; see Appendix B). Moreover, the meaning of the manifest one-variable effects and the two-variable effects remain the same as in Definition 4.1.1. This is also true for all following model definitions. Therefore, I will start the explication of this and all following definitions at the level of latent variables.

- The latent one-variable parameters $\left(\tau_x^X;\ \tau_y^Y\right)$ describe the univariate distributions of the latent variables. These parameters are identical to the odds comparing the geometric mean of all probabilities belonging to a particular latent category to the overall geometric mean. E.g.:

$$\tau_x^X = \frac{\sqrt[Y]{\prod_{y=1}^{Y} \pi_{x.y}^{X.Y}}}{\sqrt[X \cdot Y]{\prod_{v=1}^{X} \prod_{y=1}^{Y} \pi_{v.y}^{X.Y}}}, \tag{4.2.2}$$

or in proportions:

$$\tau_x^X = \frac{\sqrt[Y]{\prod_{y=1}^{Y} e_{x.y}}}{\sqrt[X \cdot Y]{\prod_{v=1}^{X} \prod_{y=1}^{Y} e_{v.y}}}, \tag{4.2.3}$$

with $e_{x.y}$ representing the expected latent cell frequency and $\pi_{x.y}^{X.Y}$ representing the latent cell proportion. The index v serves to count the categories of X when x already describes a particular category. If one knew the expected frequencies, the calculation in collapsed frequency tables would be straightforward.

- The latent two-variable effect $\left(\tau_{x.y}^{X.Y}\right)$ indicates the deviations of cell proportions from the prediction based on the marginal proportions in the latent bivariate sub-table:

$$\left(\tau_{x.y}^{X.Y}\right) = \frac{\pi_{x.y}^{X.Y}}{\pi_x^X \pi_y^Y}, \qquad (4.2.4)$$

4.2.1.2 Implications of the CT model

The CT model for categorical data has already been introduced by other authors (see e.g., Hagenaars, 1990, 1993). Since it will serve as a submodel in the Multitrait-Multirater models it will be shortly discussed.

The meaning of the parameters within a TMU remains perfectly the same as before. They may be used to determine the reliability and the meaning of the latent variables. Additionally, the association between the two latent constructs corresponds to a *heterotrait-monomethod* correlation sensu Campbell and Fiske (1959). In general, this correlation (association) should be rather low to indicate discriminant validity. However, there may also be category-specific co-occurrences that are higher than expected for the independence model. A special type of neuroticism may be related to a particular type of conscientiousness, for example. Statistically this can be seen in significant two-variable effects representing specific combinations of latent categories that are more likely to occur than predicted by the underlying latent one-variable effects.

If all two-variable parameters are equal to 1, all categories of the two constructs are perfectly distinct from each other, representing perfect discriminant validity between the two latent variables. In this case, the independence model will hold.

4.2.1.3 Definition of the independence CT model

The assumption of independent constructs (perfect discriminant validity) can be tested in log-linear models with latent variables. The independence CT model fits well, if the constructs are perfectly discriminant.

Definition 4.2.2

The independence Correlated Traits model

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \tag{4.2.5}$$

with $e_{\mathbf{ab}.x.y}$ as expected frequency of the manifest response pattern $\mathbf{ab}$. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the log-linear parameters linking the latent variable X to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the log-linear parameters linking the latent variable Y to its indicators and the manifest one-variable effects.

τ_x^X and τ_y^Y represent the latent one-variable effects.

The statistical meaning of the parameters remains perfectly the same as for the saturated model.

4.2.1.4 Applications of the CT and the independence CT model

The (categorical) CT model with two latent variables and multiple indicators will be illustrated by the empirical example of neuroticism and conscientiousness measured by four items per trait. Figure 4.2 depicts the CT-model. The first four indicators ("vulnerable", "sensitive", "moody", and "self-doubtful") measure neuroticism; the last four indicators ("industrious", "diligent", "dutiful", and "ambitious") measure conscientiousness.

Table 4.2.1 presents the goodness-of-fit coefficient for the CT and the independence CT model with 3 categories per latent variable.

Table 4.2.1

Goodness-of-fit coefficients of the CT and independence CT model with two three-categorical latent variables

	χ^2	$p\left(\chi^2\right)$	L^2	$p(L^2)$	*df*	AIC[1]	BIC[1]	p_{boot}	n_{bounds}
CT	8009.02	.00	1141.76	1.00	6504	–11866.24	–38985.39	.08	4
ind. CT	7938.05	.00	1138.94	.00	6508	–11877.06	–39012.89	.11	7

Note. CT: CT model; ind. CT: independence CT model; χ^2: Pearson χ^2-value; L^2: Likelihood-Ratio χ^2-value [1]AIC and BIC are based on the L-squared χ^2-value; p_{boot}: bootstrapped probability of χ^2; n_{bounds}: number of boundary values.

The two models fit the data according to the bootstrapped χ^2-value. According to the two information criteria the independence model fits better. However, it suffers from a larger number of boundary values than the CT model. For illustrative reasons the saturated CT model will be reported.

Results of the CT model. The one-variable effects of the manifest variables, the two-variable effects (links) of the manifest and latent variables, as well as the conditional response probabilities can be found in Tables 4.2.2 through 4.2.5.

Table 4.2.2
Log-linear parameters of the measurement model of the CT model; neuroticism

		one variable effect	two-variable effect		
variable	manifest categories		*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	0.39	2.56	1.05	0.37
	2	0.98	1.22	1.22	0.67
	3	2.61	0.32	0.78	4.00
B (sensitive)	1	$1.23\ 10^{-10}$	$1.47\ 10^{10}$	$7.12\ 10^{9}$	$9.51\ 10^{-21}$
	2	$7.68\ 10^{4}$	$1.72\ 10^{-5}$	$3.10\ 10^{-5}$*	$1.88\ 10^{9}$
	3	$1.05\ 10^{5}$	$3.95\ 10^{-6}$	$4.53\ 10^{-6}$	$5.59\ 10^{10}$
C (moody)	1	33.81	0.08	554.38	0.02
	2	20.10	0.03	749.38*	0.04
	3	0.00	420.77	$2.41\ 10^{-6}$	987.53
D (doubtful)	1	0.16	16.10	0.01	4.16
	2	1.95	0.29	12.52*	0.28
	3	3.23	0.21	5.35	0.87

Note. * boundary values. *ns*: categories of the latent variable for neuroticism.

Table 4.2.2 presents the log-linear parameters for the measurement model of neuroticism. The log-linear parameters are less aberrant than for the model presented in section 4.1.1—still, they suffer from boundary solutions.

The conditional response probabilities differ from those found for the model presented in Section 4.1.3 (examining neuroticism only). The conditional response probabilities for the 1st latent category change to a small degree only (compare Tables 4.1.9 and 4.2.3). The conditional response probabilities for the 2nd latent category for neuroticism also differ with respect to the results found in Section 4.1.3. Individuals belonging to this class tend to choose the 3rd manifest category for item *A*. They tend to choose the 2nd (moderately neurotic category) for items *B* and *D*. And they clearly do not choose the 3rd manifest response category for item *C*. Therefore, I still will call this class sensitive but (emotionally) stable[13] or middle class.

Table 4.2.3

Conditional response probabilities of the manifest response categories for the construct neuroticism in the CT model

		latent status		
variable	manifest categories	*ns* = 1	*ns* = 2	*ns* = 3
	1	.33	.11	.01
A (vulnerable)	2	.39	.33	.06
	3	.28	.56	.93
	1	.51	.23	.00*
B (sensitive)	2	.37	.64	.02
	3	.12	.13	.98
	1	.68	.55	.25
C (moody)	2	.16	.45*	.28
	3	.16	.00	.47
	1	.67	.00*	.16
D (doubtful)	2	.15	.59*	.13
	3	.18	.41	.70

Note. * boundary values. *ns*: categories of the latent variable for neuroticism.

Individuals belonging to the 3rd latent class choose the 3rd manifest response category almost with certainty for items *A* and *B*. The conditional response probability to choose the 3rd manifest response category for item *D* is less pronounced than in Table 4.1.9 but still very high (.70). Members of the 3rd class indicate that they are moody in about half of the time and tend to choose the 1st or 2nd manifest response category approximately equally often.

The latent proportions differ between the two models. In the previously described model, about one quarter of all individuals was classified as not neurotic. Approximately

[13] The name is only given for illustrative reasons.

the same amount of individuals is classified as not neurotic in the current application $\left(\hat{\pi}_1^{NEUS}=.21\right)$. Yet, the class proportions for the 2nd and 3rd class differ vastly between the models. Only 11% of the individuals are classified as sensitive but stable (middle category) – compared to 56% in Section 4.1.3. And 69% of all individuals are classified as neurotic – compared to 20% in the model of Section 4.1.3. Considering the conditional response probabilities again shows that the typical response patterns for the two applications differ in such a way that many individuals who have been classified into the 2nd category in the 1st application now belong to the third latent category. The conditional response probabilities to choose the 3rd manifest response category for items *C* and *D* became lower; but, still, it is highest compared to the other categories.

Table 4.2.4
Log-linear parameters of the measurement model of the CT model; conscientiousness

		one variable effect	two-variable effect		
variable	manifest categories		*cs* = 1	*cs* = 2	*cs* = 3
	1	0.40	9.76	0.30	0.34
E (industrious)	2	1.55	0.62	3.83	0.42
	3	1.62	0.16	0.87	6.99
	1	$1.79\ 10^{-34}$	$4.01\ 10^{34}$	$2.72\ 10^{33}$	$9.16\ 10^{-69}$*
F (diligent)	2	$8.84\ 10^{16}$	$1.04\ 10^{-17}$	$4.96\ 10^{-17}$*	$1.93\ 10^{33}$
	3	$6.33\ 10^{16}$	$2.39\ 10^{-18}$	$7.41\ 10^{-18}$	$5.65\ 10^{34}$
	1	0.22	2.95	0.57	0.59
G (dutiful)	2	1.05	1.00	1.64	0.61
	3	4.32	0.34	1.07	2.76
	1	0.86	3.85	0.76	0.34
H (ambitious)	2	1.19	1.04	1.68	0.57
	3	0.97	0.25	0.78	5.15

Note. * boundary values. *cs*: categories of the latent variable for conscientiousness.

Table 4.2.4 presents the log-linear parameters for the measurement model of conscientiousness. Inspecting the log-linear effects (but those for item *F*) reveals that the two-variable effects in one row (for one manifest category) are always strongest for those categories with identical labels of the manifest and latent category. The conditional response probabilities are highest for categories sharing the same label indicating that the 1st latent class consists of not conscientious individuals, the 2nd latent class consists of moderately conscientious individuals, and the 3rd latent class consists of highly conscientious individuals.

The only item that does not perfectly match this pattern is dutiful. This finding could be explained by the fact that the German item “pflichtbewusst” (“dutiful”) is the only item measuring conscientiousness which is both internally and externally oriented. This characteristic may stem from an internal desire to be responsible. However, it may also occur because a person is responding to strong external pressures to perform prescribed behaviors. All other adjectives describe aspects of conscientiousness that are more strongly due to attitudes (internally oriented). Therefore, it may be much easier to be dutiful or to perceive oneself as dutiful yielding principally high answers on this item, but it still fits into the latent typology. This interpretation is supported by the fact that the log-linear two-variable parameters are always highest for the categories sharing the same label (see above) and by the increase in the conditional response probabilities for higher manifest response categories combined with higher latent categories (see Table 4.2.5).

Table 4.2.5

Conditional response probabilities of the manifest response categories for the construct conscientiousness in the CT model

		latent status		
variable	manifest categories	*cs* = 1	*cs* = 2	*cs* = 3
E (industrious)	1	.76	.02	.01
	2	.19	.80	.05
	3	.05	.19	.94
F (diligent)	1	.87	.09	.00*
	2	.11	.82	.05
	3	.02	.09	.95
G (dutiful)	1	.20	.02	.01
	2	.33	.27	.05
	3	.46	.71	.94
H (ambitious)	1	.69	.19	.05
	2	.26	.59	.11
	3	.05	.22	.84

Note. * boundary values. *cs*: categories of the latent variable for conscientiousness.

Table 4.2.6 presents the latent joint distributions of the categorical traits neuroticism and conscientiousness. The marginals for neuroticism differ vastly from the previously reported model as described above. 24% of the sample are classified as not conscientious, 36% as moderately conscientious, and 41% as highly conscientious.

Table 4.2.6

Cross classification of the estimated proportions of the two latent variables (neuroticism and conscientiousness) in the latent saturated CT model

		NEUS (Neuroticism)			
		1	2	3	
CONS (Conscientiousness)	1	.05 (.05)	.02 (.03)	.17 (.17)	.24
	2	.07 (.08)	.05 (.04)	.24 (.25)	.36
	3	.09 (.09)	.04 (.05)	.28 (.28)	.41
		.21	.11	.69	

Note. Values in parentheses present the product of the two latent marginals.

Examining the joint distribution of the latent saturated CT model reveals some interesting results. The integration of the latent two-variable effect does not lead to great differences in the latent joint distribution compared to the expectations given only the latent marginals. A comparison of the estimated proportions (cell entries in Table 4.2.6) with the expected proportions given the latent one-variable effects only (in parentheses) reveals that the latent association is not very strong. This finding is supported by the estimated two-variable effects. The parameter values range from 0.80 for the latent cell combination [2 1] (brackets indicate latent cell combinations) to 1.23 for the cell combination [2 2]. The more parsimonious and better fitting independence model seems to be the model of choice for this data situation.

5 Latent Rater Agreement Models

The models presented in Section 4 serve as the basis for the adoption of manifest rater agreement models. The log-linear model with one latent variable represents a Trait-Method-Unit (TMU) in all models that will be defined. The Correlated Traits model for categorical data statistically corresponds to a latent rater agreement model if the different trait-variables are replaced by two variables representing the same trait rated by two distinct raters.

In this section, latent rater agreement models will be defined for structurally different and interchangeable raters. Structurally different raters are raters who differ from each other by the research design. Consider self- and peer ratings as a typical example. The self-raters can be randomly drawn out of the population of all available self-raters. The peers can then be drawn out of the set of possible peer raters. Self- and peer raters stem from different populations and are, therefore, structurally different.

The opposite accounts for interchangeable raters. Drawing two peers out of the set of possible peer raters corresponds to random sampling out of one population. Random samples of one population must have the same parameters. Therefore, the models for interchangeable raters are restricted versions of the models for structurally different raters.

5.1 Latent Rater Agreement Models for Structurally Different Raters

The definition of latent rater agreement models is based on the previously described log-linear models with latent variables. However, the two distinct construct of the Correlated Traits (CT) model are replaced by two variables representing one construct rated by a self- and a peer rater. The structure of the model remains perfectly the same (see Figure 5.1).

The latent rater agreement models allow for a test if the latent categories represent the same latent constructs. If this is the case, the two ratings must principally be classifiable into the same number of categories with identical labels.

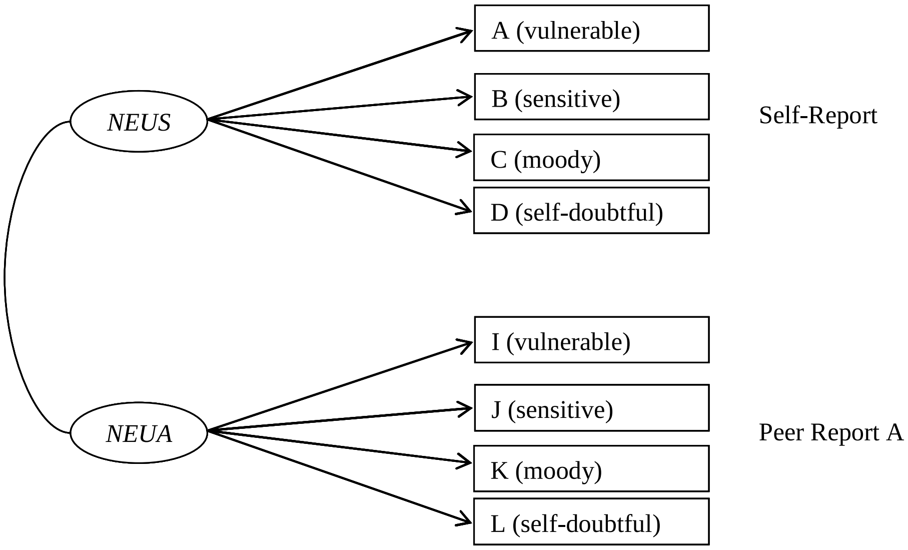

Figure 5.1. Log-linear model with two latent variables representing the latent construct Neuroticism (*NEU*) for the self-report *S* and the peer report *A*. Each latent variable is measured by four manifest indicators.

Figure 5.1 presents a categorical monotrait-multimethod model for the analysis of latent rater agreement of two raters. For sake of comprehensibility the trait variables and the items are labeled. The latent construct neuroticism is represented by two latent variables (class variables: *NEUS* for the self-report vs. *NEUA* for the peer rating A). The two latent variables are measured by the same set of items (“vulnerable”, “sensitive”, “moody”, and “self-doubtful”) rated by a self-rater and one peer. However, administering the same items is not a necessary condition for the definition and application of the models.

Out of the total of four existing manifest rater agreement models (see Section 2.3) three models can be chosen to analyze rater agreement at the latent level for structurally different raters. The quasi-independence I model (5.1.1), the quasi-independence II model (5.1.2), and the quasi-symmetry model (5.1.3) can be defined for structurally different raters. The symmetry model implies interchangeability of the raters and will be presented in Section 5.2. The independence model and the saturated model have been defined in Section 4.1. The definitions apply directly to the case of two methods measuring the same trait. All models will be defined for the case of two raters.

In all models that will be presented, there are two coefficients that may be determined revealing information about bias and distinguishability. Since the latent one-variable effects do not always directly reflect the univariate latent distributions of the latent variables, the coefficients are defined relying on the latent probabilities. Differences in the prevalence rates (differences in the latent distributions) represent method (rater) bias (see Agresti, 1992). The method bias type I coefficient quantifies this effect.

Definition 5.1.1

Method bias type I

$$MB1_{(X.Y)} = \frac{\pi_x^X}{\pi_y^Y}\text{, for } x = y \tag{5.1.1}$$

is the method bias type I coefficient.

This definition of method bias is similar to the conception of method bias in standard log-linear models for rater agreement (see Agresti, 1992). Note, that this bias is not defined as a bias indicating differences / deviations from the true status or the true distribution of the latent variable but as a bias with respect to the other rater. Values larger than 1 indicate that the rater whose latent variable is in the numerator uses this category more frequently than the other rater. Values below 1 indicate the opposite. High (or low) values on *MB*1 indicate that the two raters do not perfectly agree on the prevalence rates and therefore also indicates a cause of a lack of convergent validity.

In all models, a second type of bias can be examined. The ratio to which proportions of specific cell-combinations besides the main diagonal deviate from the expected proportions given the one-variable effects is defined as *distinguishability index*. This index is a direct consequence of the concept of distinguishable categories formulated in Section 2.3.1. To my knowledge it has not been defined yet.

Definition 5.1.2

Distinguishability index (Dist)

$$Dist_{(x.y)} = \frac{\pi_{x.y}^{X.Y}}{\pi_x^X \pi_y^Y}\text{, for } x \neq y\,. \tag{5.1.2}$$

The distinguishability index indicates to which ratio particular cells of the joint distribution representing discordant ratings are over- or underrepresented. Values larger than 1 indicate that the proportions of the cell combinations *x.y* is higher than expected given the latent marginals, values smaller than 1 indicate that these proportions are smaller than expected given the latent marginals. If the values are larger than 1, the two raters confound the categories *x* and *y*. That is, if category *x* is chosen the probability to observe category *y* increases; the two raters do not appropriately distinguish between these two categories. If the index is smaller than 1 the two raters produce smaller latent proportions for these cells than expected given the marginals and, therefore, they distinguish between these categories—the closer this value is to 0, the better the raters distinguish between the two particular categories. A further analysis inspecting the moderators of agreement (see Funder, 1995) could reveal why raters confound or distinguish well between different categories.

If raters distinguish perfectly between all categories they also agree perfectly implying that a one-variable model will hold. The one-variable model can be defined as specified in Equation 4.1.1 (where all items depend uniquely on one common latent variable).

5.1.1 Definition of the Quasi-Independence I Model for Structurally Different Raters

Definition 5.1.3

The latent quasi-independence I model for two structurally different raters and one construct

Let *X* and *Y* represent the same latent construct measured by two distinct raters with identical categories (*x* and *y*).

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \left(\tau_{x.y}^{X.Y}\right)^I, \quad \text{with} \begin{cases} I = 1 \text{ if } x = y \\ I = 0 \text{ else} \end{cases} \qquad (5.1.3)$$

$e_{\mathbf{ab}.ns.cs}$ is the expected frequency of a specific cell in the latent joint cross-classification of the manifest response patterns $\mathbf{ab}$ (consisting of the two trait-specific patterns $\mathbf{a}$ and $\mathbf{b}$) with the two latent variables *X* and *Y*. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the product of the log-linear parameters linking the latent variable *X* to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the product of the log-linear parameters linking the latent variable *Y* to its indicators and the manifest one-variable effects.

τ_x^X and τ_y^Y represent the latent one-variable effects. $\tau_{x.y}^{X.Y}$ represents the latent two-variable effects.

5.1.1.1 The statistical meaning of the different effects in the latent quasi-independence I model for structurally different raters

The log-linear parameters of Definition 5.1.3 have the following meanings:

- η is the geometric mean of the unobserved complete frequency table, which is a mere reflection of the sample size (Hagenaars, 1990; 1993).

- The submodels $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$: have been described in section 4.1 (e.g., Goodman, 1974a, 1974b; Haberman, 1979; Hagenaars, 1990, 1993; McCutcheon, 1987).

- The latent one-variable parameters $\left(\tau_x^X ; \tau_y^Y\right)$ cannot be interpreted as in the models described before. As for the manifest quasi-independence models the table of expected proportions can be decomposed into one table showing perfect agreement (a one-variable model holds) and one part following complete independence. For the model with manifest variables only, Schuster and Smith (2006) showed that the cell proportions in the part with perfect agreement (that is, a part of the cells on the main diagonal) only depend on the additional log-linear parameters $\left(\tau_{x.y}^{X.Y}\right)$ and that the one-variable effects account for the remaining part. Adapted Equation 4 from Schuster and Smith (2006) is:

$$\pi_{x.y}^{X.Y} = \frac{\eta\left[\left(\tau_{x.y}^{X.Y}\right)^{I}\tau_x^X\tau_y^Y\right]}{N}, \tag{5.1.4}$$

with *N* indicating the sample size. For cases when $x \neq y$, this simplifies to:

$$\begin{aligned}\pi_{x.y}^{X.Y} &= \frac{\eta\left[\left(\tau_{x.y}^{X.Y}\right)^{0}\tau_x^X\tau_y^Y\right]}{N}\\ &= \frac{\eta\left[1\times\tau_x^X\tau_y^Y\right]}{N}\\ &= \frac{\eta\tau_x^X\tau_y^Y}{N}\end{aligned} \quad . \tag{5.1.5}$$

The log-linear effects can be determined using the following equations:

$$\tau_x^X = \frac{\pi_x^X - \dfrac{\tau_{x.y}^{X.Y}}{\sum_{v=1}^{X}\left(\tau_v^X\right)\sum_{y=1}^{Y}\left(\tau_y^Y\right)+\tau_{x.y}^{X.Y}}}{\sqrt[X]{\prod_{v=1}^{X}\pi_v^X}}, \tag{5.1.6}$$

$$\tau_y^Y = \frac{\pi_y^Y - \dfrac{\tau_{x.y}^{X.Y}}{\sum_{x=1}^{X}\left(\tau_x^X\right)\sum_{w=1}^{Y}\left(\tau_y^Y\right)+\tau_{x.y}^{X.Y}}}{\sqrt[Y]{\prod_{w=1}^{Y}\pi_w^Y}}, \tag{5.1.7}$$

and

$$\tau_{x.x}^{X.Y} = \frac{\pi_{x.x}^{X.Y}}{\pi_{x.x}^{X.Y} - \dfrac{\tau_x^X\tau_x^Y}{\left(\sum_{v=1}^{X}\tau_v^X\right)\left(\sum_{w=1}^{Y}\tau_w^Y\right)}}, \tag{5.1.8}$$

with *v* and *w* indicating the categories of *X* and *Y*, respectively. Repeating the same index *x.x* instead of specifying *x.y* means that *y* = *x*. As can be seen the log-linear parameters can be determined knowing the latent proportions. Additionally, the parameters always can be determined relying on the decomposition of the latent joint distribution as proposed by Schuster and Smith (2006) for manifest variables. They conceive the joint manifest distribution as a mixture of ambiguous and obvious cases. Ambiguous cases are target persons upon whom the two raters do not agree or only due to chance agreement. This rationale can directly be adopted at the latent level. For ambiguous targets the independence model holds:

$$\tau^{\circ X}_{x} = \frac{\pi^{\circ X}_{x}}{\sqrt[X]{\prod_{v=1}^{X} \pi^{\circ X}_{v}}}, \tag{5.1.9}$$

with ° marking that only the ambiguous cases are concerned. In order to obtain the marginals of the latent table following independence, the amount of overrepresentation on the main diagonal has to be subtracted. This is done in Equations 5.1.6 and 5.1.7. Equation 5.1.8 may then be used to determine the latent two-variable effect. However, these parameters are not directly related to the proportions; therefore one typically relies on the expected proportions reporting the quasi-independence models.

5.1.1.2 Implications of the quasi-independence I model

The latent one-variable effects do not directly reflect the univariate latent distributions of the latent variables. Their interpretation is rather difficult with respect to the complete table, but much easier with respect to the decomposed table (separating ambiguous from obvious cases). The method bias type I can be determined revealing differences between the latent prevalence rates (latent distributions).

Concordant ratings (agreement) which go beyond the agreement on chance are indicated by the two-variable effects $\left(\tau^{X.Y}_{x.y}\right)^{1}$ for cells with identical indices $\left(x = y\right)$. Agreement for raters is a special case of convergence in general. Thus, these parameters depict the category-specific convergence beyond chance convergence. An overall latent agreement rate can be calculated using κ. A category-specific agreement rate can be calculated by the ratio of the expected cell proportion to the product of the latent marginals. Large differences in the category-specific agreement rates indicate that raters agree more or less strongly depending on the categories of the latent variables. Large differences indicate that the convergent validity (agreement) depends on the categories and is not constant across categories.

By fitting the latent quasi-independence I model the assumption of independent disagreement is tested. Therefore, it is not meaningful to calculate the distinguishability index in quasi-independence models.

5.1.2 Definition of the Quasi-Independence II Model for Structurally Different Raters

Definition 5.1.4

The latent quasi-independence II model

Let *X* and *Y* represent the same latent construct measured by two distinct raters with identical categories (*x* and *y*).

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \left(\tau_{x.y}^{X.Y}\right)^I, \text{ with } \begin{cases} I = 1 \text{ if } x = y \\ I = 0 \text{ else} \end{cases}, \text{ and } \tau_{x.y}^{X.Y} = \tau^{X.Y}. \tag{5.1.10}$$

$e_{\mathbf{ab}.ns.cs}$ is the expected frequency of a specific cell in the latent joint cross-classification of the manifest response patterns $\mathbf{ab}$ (consisting of the two trait-specific patterns $\mathbf{a}$ and $\mathbf{b}$) with the two latent variables *X* and *Y*. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the product of the log-linear parameters linking the latent variable *X* to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the product of the log-linear parameters linking the latent variable *Y* to its indicators and the manifest one-variable effects.

τ_x^X and τ_y^Y represent the latent one-variable effects. $\tau_{x.y}^{X.Y}$ represents the latent two-variable effects. It is restricted to be constant across all cells on the main diagonal $\left(\tau_{x.y}^{X.Y} = \tau^{X.Y}\right)$.

The statistical meaning of the parameters is absolutely identical to the meaning of the parameters of the quasi-independence I model.

5.1.2.1 Implications of the quasi-independence II model

Concordant ratings which go beyond the agreement on chance are mirrored by the two-variable effects $\left[\left(\tau^{X.Y}\right)^1\right]$ for cells with identical indices $\left(x = y\right)$. These effects show *constant* agreement between raters. Agreement is a property of the raters and not of the interaction between raters and categories. The two-variable effects in manifest models can be transformed into κ (see Schuster & Smith, 2006). The two-variable parameters depict the convergence beyond chance convergence. An overall latent agreement rate can be calculated using κ. A category-specific agreement rate can be calculated by the ratio of the expected cell proportion to the prediction given the marginals only.

By fitting the latent quasi-independence II model the assumption of constant independent disagreement is tested. Therefore, it is not meaningful to calculate the distinguishability index in quasi-independence models.

5.1.3 Definition of the Quasi-Symmetry Model for Structurally Different Raters

Definition 5.1.5

The latent quasi-symmetry model

Let X and Y represent the same latent construct measured by two distinct raters with identical categories (x and y).

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \tau_{x.y}^{X.Y}\,, \qquad \text{with } \tau_{x.y}^{X.Y} = \tau_{y.x}^{X.Y} \tag{5.1.11}$$

$e_{\mathbf{ab}.ns.cs}$ is the expected frequency of a specific cell in the latent joint cross-classification of the manifest response patterns $\mathbf{ab}$ (consisting of the two trait-specific patterns $\mathbf{a}$ and $\mathbf{b}$) with the two latent variables X and Y. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the log-linear parameters linking the latent variable X to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the log-linear parameters linking the latent variable Y to its indicators and the manifest one-variable effects.

τ_x^X and τ_y^Y represent the latent one-variable effects. $\tau_{x.y}^{X.Y}$ represents the latent two-variable effects.

5.1.3.1 The statistical meaning of the different effects in the latent quasi-symmetry model for structurally different raters

The meaning of the log-linear parameters of Definition 5.1.11 directly corresponds to the log-linear parameters of the saturated model, however, some restrictions are imposed:

- η is the geometric mean of the unobserved complete frequency table.

- The submodels $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$: have been described in section 4.1 (e.g., Goodman, 1974a, 1974b; Haberman, 1979; Hagenaars, 1990, 1993; McCutcheon, 1987).

- The latent one-variable parameters τ_x^X and τ_y^Y: describe the univariate distributions of the latent variables. These parameters are identical to the odds comparing the geometric mean of a particular category (*x* or *y*) against the geometric mean of all cells. E.g.:

$$\tau_x^X = \frac{\sqrt[Y]{\prod_{y=1}^{Y} \pi_{x.y}^{X.Y}}}{\sqrt[X\cdot Y]{\prod_{v=1}^{X}\prod_{y=1}^{Y} \pi_{v.y}^{X.Y}}}, \qquad (5.1.12)$$

  with *x* indicating the particular latent category of *X* and *v* indexing the first to the last category of *X* in the denominator.

- The latent two-variable effect $\left(\tau_{x.y}^{X.Y}\right)$ indicates the deviations of joint cell proportions from the prediction based on the marginal proportions in the latent bivariate sub-table. E.g.:

$$\left(\tau_{x.y}^{X.Y}\right) = \left(\tau_{y.x}^{X.Y}\right) = \sqrt{\frac{\pi_{x.y}^{X.Y}\pi_{y.x}^{X.Y}}{\pi_x^X \pi_y^Y \pi_y^X \pi_x^Y}},$$

$$\text{(5.1.13)}$$

$$\text{simplifying to } \left(\tau_{x.y}^{X.Y}\right) = \frac{\pi_{x.y}^{X.Y}}{\pi_x^X \pi_y^Y} \text{ for } x = y\,.$$

5.1.3.2 Implications of the quasi-symmetry model

The latent one-variable effects reflect the univariate latent distributions of the latent variables. Differences between the latent distributions originate in different (perceived) prevalence rates; therefore, differences in the latent distributions represent method bias (see Agresti, 1992).

The distinguishability index can be used to analyze the ratio to which the expected proportions of a disagreement cell deviate from the product of the marginal expected proportions. It is the (geometric) mean of the over- or underrepresentation of specific disagreement cells. In the quasi-symmetry model, the over- or underrepresentation by definition follows a specific pattern of interchangeability: the two-variable effects are restricted to be equal for any pair of categories that consists of the same categories $\left[\left(\tau_{x.y}^{X.Y}\right) = \left(\tau_{y.x}^{X.Y}\right)\right]$. However, this does not necessarily afflict the distinguishability index except for the case of identical latent marginals:

$$Dist_{(x.y)} = \frac{\pi_{x.y}^{X.Y}}{\pi_x^X \pi_y^Y} = \frac{\pi_{y.x}^{X.Y}}{\pi_y^X \pi_x^Y} = \left(\tau_{x.y}^{X.Y}\right) = \left(\tau_{y.x}^{X.Y}\right) = Dist_{(y.x)}\,, \qquad (5.1.14)$$

only if $\pi_x^X = \pi_x^Y$ and $\pi_y^X = \pi_y^Y$.

This implies that the quasi-symmetry model may be used to test if the underlying pattern of disagreement follows a symmetric structure but it does not test if the ratio of over- or underrepresentation as examined by the distinguishability index is the same. This can be done applying the symmetry model (which will be presented for the case of interchangeable raters).

Concordant ratings which go beyond the agreement on chance are reflected by the two-variable effects on the main diagonal $\left(\tau_{x.y}^{X.Y}\right)$ for $x = y$. These effects show agreement between raters. Agreement between raters is a special case of convergence in general. Thus, these parameters depict the category-specific convergence beyond chance convergence. An overall latent agreement rate can be calculated using κ. A category-specific agreement rate can be calculated by the ratio of the probability of a cell combination representing agreement to the product of the marginals.

5.1.4 Applications of the Latent Rater Agreement Models for Structurally Different Raters

The latent rater agreement models for structurally different raters and multiple indicators will be illustrated by the empirical example of neuroticism measured by the self-report and the first peer report (peer *A*). The data have been described in Section 4.1.3.

Table 5.1.1

Goodness-of-fit coefficients of the rater agreement models with three-categorical variables for structurally different raters

	χ^2	$p\left(\chi^2\right)$	L^2	$p(L^2)$	*df*	AIC[1]	BIC[1]	p_{boot}	n_{bounds}
sat	7935.28	.00	1464.35	1.00	6504	–11543.65	–38662.80	.23	8
ind	7768.37	.00	1496.54	1.00	6508	–11519.46	–38655.29	.18	10
QI-I	7897.34	.00	1466.62	1.00	6505	–11543.38	–38666.70	.15	8
QI-II	8061.78	.00	1469.85	1.00	6507	–11544.15	–38675.81	.15	5
QS	7897.26	.00	1466.62	1.00	6503	–11539.38	–38654.36	.16	10
ONE	7880.61	.00	1518.76	1.00	6510	-11501.24	-38645.40	.22	2

Note. sat: saturated model; ind: independence model; QI-I; quasi-independence I model; QI-II: quasi-independence II model; QS: quasi-symmetry model; ONE: one-variable model; χ^2 : Pearson χ^2 -value; L^2: Likelihood-Ratio χ^2 -value; [1]AIC and BIC are based on the L-squared χ^2 -value; p_{boot}: bootstrapped probability of χ^2 ; n_{bounds}: number of boundary values.

Figure 5.1 depicts the saturated model. The first four indicators (“vulnerable”, “sensitive”, “moody”, and “self-doubtful”) measure neuroticism (*NEUS*) in the self-report form; and the (identically worded) last four indicators measure neuroticism (*NEUA*) in the peer-report form (for peer A).

The empirical χ^2-values (presented in Table 5.1.1) do not approximate their theoretical distributions (very different probabilities associated to these values for the Pearson and likelihood-based coefficients). Therefore, one should rely on the bootstrap analysis to identify models that fit to the data. According to the bootstrap analyses all models fit to the data. Inspecting the information criteria (AIC and BIC) reveals that the quasi-independence II latent rater agreement model fits best.

The saturated latent rater agreement model. The saturated rater agreement model has not explicitly been defined in this section. However, its definition is absolutely identical to the CT model presented in Section 4.1. It fits to the data with respect to the bootstrapped χ^2-value. 8 log-linear parameters suffer from boundary values. The one- and two-variable effects related to the manifest variables can be found in Appendix C.

Table 5.1.2 presents the conditional response probabilities for neuroticism in the self-report. These do virtually not differ from the values presented in Table 4.1.9. Therefore, their values will not be interpreted here.

Table 5.1.2

Conditional response probabilities of the manifest response categories for the construct neuroticism (NEUS) in the saturated latent rater agreement model for structurally different raters (self-report)

variable	manifest categories	latent status *ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	.31	.03	.00
	2	.43	.11	.00
	3	.26	.86	1.00
B (sensitive)	1	.51	.03	.00*
	2	.46	.10	.00*
	3	.03	.87	1.00*
C (moody)	1	.68	.38	.06
	2	.20	.40	.07
	3	.12	.22	.88
D (doubtful)	1	.51	.25	.01
	2	.31	.16	.12
	3	.18	.59	.87

Note. * boundary values. *ns*: categories of *NEUS*.

Table 5.1.3 provides the conditional response probabilities for peer *A*. The peers may also be divided into three latent classes showing different typical response patterns. Individuals of the first class (20%) clearly favor the first response category across all items. The conditional response probabilities for the 1st manifest category are much higher than for the self-report.

Individuals belonging to the 2nd latent class (51%) show typical response patterns that are spread across all possible response categories. The highest conditional response probability for this class is $\left(\pi_{1.2}^{C.NEUA} = .57\right)$ to choose the 1st category of rating the target to be moody, and the lowest conditional response probability for this class is $\left(\pi_{1.2}^{A.NEUA} = .08\right)$ to choose the 1st category of vulnerable with respect to the target. 7 out of 12 conditional response probabilities are in the range between $\left(\pi_{2.2}^{D.NEUA} = .32\right)$ and $\left(\pi_{3.2}^{A.NEUA} = .49\right)$ illustrating that this class of individuals uses all manifest response categories.

Table 5.1.3

Conditional response probabilities of the manifest response categories for the construct neuroticism (NEUA) in the saturated latent rater agreement model for structurally different raters (peer report A)

variable	manifest categories	latent status *na* = 1	 *na* = 2	 *na* = 3
	1	.62	.08	.00*
I (vulnerable)	2	.38	.44	.00*
	3	.00*	.49	1.00*
	1	.74	.18	.00*
J (sensitive)	2	.26	.47	.09
	3	.00*	.35	.91
	1	.76	.57	.34
K (moody)	2	.15	.31	.25
	3	.09	.11	.41
	1	.76	.48	.21
L (doubtful)	2	.14	.32	.26
	3	.10	.20	.53

Note. * boundary values; *na*: categories of *NEUA*.

Individuals belonging to the 3rd latent class (29%) interestingly show a typical response pattern which is similar to the typical response pattern of the 2nd latent class of the self-raters. That is, these individuals clearly rate the target to be vulnerable and sensitive but they have no very pronounced view about the target's moodiness and self-doubts. However, the conditional response probabilities to choose the 3rd manifest response categories are highest for this latent class.

The two raters differ with respect to their measurement models. However, the interpretations of their conditional response probabilities are close to each other. Both types of raters can be classified in ordered categories. Their measurement models differ with respect to the difficulty of the items but not in the patterns. The assumption of measurement equivalence could be tested restricting the log-linear parameters linking the latent to the manifest variables.

Table 5.1.4

Cross classification of the two latent variables (NEUS and NEUA) in the saturated latent rater agreement model for structurally different raters

			NEUA		
		1	2	3	
	1	.09 (.05) [1.96]	.12 (.12) [1.21]	.02 (.07) [0.42]	.23 [0.73]
NEUS	2	.07 (.11) [0.68]	.32 (.27) [1.36]	.14 (.15) [1.08]	.53 [1.73]
	3	.04 (.05) [0.75]	.07 (.12) [0.61]	.13 (.07) [2.18]	.24 [0.80]
		.20 [0.71]	.51 [1.58]	.29 [0.90]	

Note. *NEUS*: neuroticism self-rating; *NEUA*: neuroticism peer rating (A); the product of the marginals is presented in parentheses; log-linear parameters are presented in brackets.

Table 5.1.4 presents the cross-classification of the latent categories of the self- and peer ratings with respect to neuroticism. The marginals present the proportions of individuals in the sample belonging to the three classes of self- or peer-rated neuroticism. The latent distributions do virtually not differ from each other. This can be seen by inspecting the latent marginals and / or the corresponding log-linear parameters (in brackets) and this proofs true calculating method bias type I:

$$\begin{aligned} MB1_{(ns=1.na=1)} &= \frac{.23}{.20} = 1.15 \\ MB1_{(ns=2.na=2)} &= \frac{.53}{.50} = 1.06\,, \\ MB1_{(ns=3.na=3)} &= \frac{.24}{.29} = 0.83 \end{aligned} \tag{5.1.15}$$

The two latent distributions do not differ strongly from each other as is indicated by *MB*1 coefficients close to 1. The hypothesis that the two raters (self and peer *A*) produce identical latent proportions could be tested in a model with restricted latent one-variable parameters (see also Section 5.2 for interchangeable raters).

Inspecting the latent joint distribution reveals that cells on the main diagonal are much more frequently expected than cells besides the main diagonal (a total of 54% entries are on the main diagonal; all three two-variable parameters are larger than 1). Moreover, comparing the expected cell frequencies with the frequencies one would expect given independent ratings (values in parentheses; assuming all other effects to be equal in this model) reveals that only cells on the main diagonal are more often observed than predicted by their corresponding margins. $\kappa = .25$ indicates low agreement between the two raters. The ratios of expected cell proportion and cell proportion based on the marginals are 1.8 for cell [1 1], 1.19 for cell [2 2], and 1.86 for cell [3 3]. That is, agreement is much higher for the 1st and 3rd class of neuroticism than for the middle category.

Table 5.1.5

Distinguishability indices for the saturated latent rater agreement model for structurally different raters

		NEUA		
		1	2	3
	1		1.00	0.29
NEUS	2	0.64		0.93
	3	0.80	0.58	

Note. *NEUS*: neuroticism self-rating; *NEUA*: neuroticism peer rating (A).

The distinguishability indices (see Table 5.1.5) show an interesting pattern. Self-raters and peers distinguish well between the extreme category combinations. That is, the combinations [1 3] and [3 1]. They also distinguish well between the middle category for the self-rating and the lowest category for the peer rating [2 1] as well as between the highest category for the self-rating and the middle category for the peer rating [3 2]. They do not distinguish (but also do not confound) the category combinations [1 2] and [2 3]. If the self-rating is considered as a gold-standard one may conclude that the peer rarely underestimates the latent score (respecting the ordered structure of the latent classes). The peer rarely extremely overestimate the latent score (choosing category 3 when the self-rating is lowest), but overestimates the latent score for the lowest and middle category.

Given the interesting similarity between the 2nd class of the self-raters and the 3rd class of the peer-raters one might think that a labeling problem occurred and that these two classes consist of sensitive but stable individuals. This is not the case. The latent joint distribution clearly shows that there is no overrepresentation compared to chance effects for the latent cells [2 3] and [3 2], which would indicate a shift in the labels, but there is an overrepresentation for the latent cells [2 2] and [3 3] representing similar classifications. One may speculate that peers are able to detect if a friend is vulnerable and sensitive but that they do not perceive the moodiness and the self-doubts of their friends as their friends do not frankly present them in their behavior. Intuitively, this is very reasonable because individuals being in a bad mood or being in a phase of severe self-doubts may not search for their friends' company and, therefore, their friends cannot comment on these items with certainty. This finding also fits well to the aspect of availability in the realistic accuracy model (see Funder, 1995).

The independence model. This model fits to the data with respect to the bootstrapped χ^2-value but it fits second worst to the data according to the AIC and BIC indices. 10 log-linear parameters suffer from boundary values. The parameters of the measurement models do not change compared to those of the saturated model. Therefore, their interpretation is perfectly the same.

Table 5.1.6 presents the latent joint distribution for the independence model. The latent marginals for the self-report do virtually not differ from the saturated model. The marginals for the peer report differ slightly from those previously reported. The 2nd category is less frequently expected than in the saturated model and the 3rd category is more frequently expected.

Table 5.1.6

Cross classification of the two latent variables (NEUS and NEUA) in the independence model for structurally different raters

		NEUA			
		1	2	3	
NEUS	1	.06	.10	.08	.24 [0.80]
	2	.13	.24	.19	.56 [1.88]
	3	.05	.09	.07	.21 [0.66]
		.24 [0.72]	.43 [1.33]	.34 [1.04]	

Note. *NEUS*: neuroticism self-rating; *NEUA*: neuroticism peer rating (A); log-linear parameters are presented in brackets.

The two raters differ more strongly from each other than in the saturated model. There is no method bias type I for the 1st category (the two categories are expected with equal proportions), however, the expected proportions for the 2nd and 3rd category differ to a greater extend. Self-raters belong more often to the 2nd latent class whereas peer ratings tend to belong to the 3rd latent class more often:

$$\begin{aligned} MB1_{(ns=1.na=1)} &= \frac{.24}{.24} = 1.00 \\ MB1_{(ns=2.na=2)} &= \frac{.56}{.43} = 1.30\,, \\ MB1_{(ns=3.na=3)} &= \frac{.21}{.34} = 0.62 \end{aligned} \tag{5.1.16}$$

A calculation of the distinguishability index is not meaningful since the latent table follows the assumption of independence.

The quasi-independence I latent rater agreement model. The quasi-independence I latent rater agreement model fits to the data according to the bootstrap results. Additionally, it fits 2nd best to the data according to the information criteria. Again, 8 log-linear parameters suffer from boundary values. The conditional response probabilities are almost identical to the conditional response probabilities reported before.

Table 5.1.7 presents the latent joint distribution of the latent categories of the self- and peer ratings. Compared to the latent proportions found for the self-report in Table 4.1.9 the 2nd latent category is underrepresented and the 3rd latent category is overrepresented. However, these differences are not very large. This may be due to the fact that the conditional response probability to choose the 3rd category for moody is somewhat lower in this application than in the application for the self-ratings only (see Table 4.1.9). Therefore, more self-raters provide response patterns which fit into this category. The log-linear parameters (presented in brackets) cannot be directly related to the latent proportions. Therefore, it is much more convenient to analyze the latent proportions. Virtually the two latent marginal distributions do not differ from each other. Inspecting the method bias type I reveals a very similar picture:

$$\begin{aligned} MB1_{(ns=1.na=1)} &= \frac{.23}{.20} = 1.15 \\ MB1_{(ns=2.na=2)} &= \frac{.50}{.49} = 1.02\ . \\ MB1_{(ns=3.na=3)} &= \frac{.28}{.32} = 0.88 \end{aligned} \qquad (5.1.17)$$

Table 5.1.7

Cross classification of the two latent variables (NEUS and NEUA) in the quasi-independence I latent rater agreement model for structurally different raters

		NEUA			
		1	2	3	
NEUS	1	.09 (.05) [3.42]	.10 (.11)	.04 (.07)	.23 [0.69]
	2	.09 (.10)	.29 (.25) [0.87]	.12 (.16)	.50 [2.24]
	3	.02 (.06)	.10 (.03)	.16 (.09) [4.62]	.28 [0.65]
		.20 [0.57]	.49 [2.25]	.32 [0.78]	

Note. *NEUS*: neuroticism self-rating; *NEUA*: neuroticism peer rating (A); the product of the marginals is presented in parentheses; log-linear parameters are presented in brackets.

Inspecting the latent joint distribution reveals that cells on the main diagonal are much more frequently expected than cells besides the main diagonal (a total of 54% entries are on the main diagonal). Moreover, comparing the expected cell frequencies with the frequencies one would expect given independent ratings (values in parentheses; assuming all other effects to be equal in this model) reveals that all cells on the main diagonal are more often observed than predicted by their corresponding margins. However, $\kappa = .18$ indicates very low agreement between the two raters. The ratios of expected cell proportion and cell proportion based on the marginals are 1.8 for cell [1 1], 1.16 for cell [2 2], and 1.78 for cell [3 3]. That is, agreement is much higher for the 1st and 3rd class of neuroticism as could be found for the saturated model. It is not meaningful to compute the distinguishability index for this model because the cell proportions of the disagreement cells follow an independence pattern.

The quasi-independence II latent rater agreement model. The quasi-independence II latent rater agreement model fits to the data according to the bootstrapped χ^2-value and it fits best to the data according to the information criteria. Five log-linear parameters suffer from boundary values. The conditional response probabilities are almost identical to the conditional response probabilities reported before.

The quasi-independence II model shows considerably differing latent marginal distributions for the latent categories of the self-report compared to those of the saturated model (see Table 5.1.4). Compared to the saturated model, class 2 is about 12% smaller and class 3 is about 12% larger. A similar - yet less strong - decline and increase can be found for the latent classes of the peers (minus 5% in the 2nd class and plus 5% in the 3rd class). This is due to the fact, that the overall agreement (the sum of all proportions on the main diagonal) is fitted in this model and not the cell-specific agreement (see e.g., Nussbeck, 2006).

The latent joint distribution shows considerable overrepresentation on the main diagonal and considerably lower expected cell proportions besides the main diagonal compared to the product of the latent marginals. Note, that the log-linear parameter indicating the overrepresentation on the main diagonal is constant.

Table 5.1.8

Cross Classification of the two latent variables (NEUS and NEUA) in the quasi-independence II latent rater agreement model for structurally different raters

		NEUA			
		1	2	3	
NEUS	1	.11 (.06) [2.53]	.08 (.12)	.06 (.09)	.25 [0.81]
	2	.06 (.09)	.27 (.19) [2.53]	.08 (.14)	.41 [1.12]
	3	.06 (.08)	.11 (.17)	.19 (.12) [2.53]	.36 [1.11]
		.23 [0.74]	.46 [1.37]	.34 [0.98]	

Note. *NEUS*: neuroticism self-rating; *NEUA*: neuroticism peer rating (A); the product of the marginals is presented in parentheses; log-linear parameters are presented in brackets.

The method bias type I coefficients for the two raters are minimal indicating that the quasi-independence II model predicts almost perfectly the same latent marginals:

$$
\begin{aligned}
MB1_{(ns=1.na=1)} &= \frac{.25}{.23} = 1.09 \\
MB1_{(ns=2.na=2)} &= \frac{.41}{.46} = 0.89\,. \\
MB1_{(ns=3.na=3)} &= \frac{.36}{.34} = 1.06
\end{aligned}
\qquad (5.1.18)
$$

Since the disagreement follows an independence pattern it is not meaningful to compute the distinguishability indices. $\kappa = .32$ indicates low agreement between the two raters. The ratios of expected cell proportion and cell proportion based on the marginals are 1.83 for cell [1 1], 1.42 for cell [2 2], and 1.58 for cell [3 3]. That is, agreement is much higher for the 1st and 3rd class of neuroticism. Although the rate of agreement depicted by the latent two-variable log-linear parameter is constant, the expected proportions on the main diagonal do not have to be overrepresented to the same ratio given the latent marginals. This is due to the fact that the log-linear parameters of the quasi-independence models do not directly relate to frequencies or proportions.

The quasi-symmetry latent rater agreement model. The quasi-symmetry latent rater agreement model fits to the data according to the bootstrapped χ^2-value, however, if fits worse than the other models according to the AIC and BIC index. This model suffers from a problem due to too many parameters (which can also be seen for the saturated model). This is in line with the increase in boundary values which indicate the problems during the estimation process. Ten log-linear parameters suffer from boundary values. As for the other models, the one- and two-variable effects related to the manifest variables as well as the conditional response probabilities can be found in Appendix C. The conditional response probabilities are almost identical to the conditional response probabilities reported before.

Inspecting the latent joint distribution reveals that the latent proportions are close to what has been found for the other models. There is a considerable overrepresentation on the main diagonal indicating agreement between the raters. Additionally, the cells besides the main diagonal follow quasi-symmetry. That is,, their two variable effects are the same for cells representing a particular combination of categories and its inversed (e.g., [1 2] and [2 1]).

Unfortunately, this model cannot be specified in LEM (Vermunt, 1997a) relying on contrast coding but has to be specified relying on dummy-coding (for a description of dummy coding see e.g., Hagenaars, 1993). Therefore, the log-linear parameters cannot be interpreted in the ways described above. The parameters are depicted in Table 5.1.9, the latent category combination [3 3] is the reference category $\left(\tau_{3.3}^{NEUS.NEUA} = 1.00\right)$. Its expected proportion can be determined by the product of the corresponding one-variable parameters. The one-variable parameters depict the (geometric) mean deviation of the corresponding rows or columns from the reference category. The two-variable effects depict the deviations from the corresponding cells from the product of the expected proportion of the reference category and the one-variable parameters.

Table 5.1.9

Cross classification of the two latent variables (NEUS and NEUA) in the quasi-symmetry latent rater agreement model for structurally different raters

		NEUA			
		1	2	3	
NEUS	1	.09 (.05) [1.04]	.10 (.11) [0.97]	.04 (.07) [0.26]	.23 [0.96]
	2	.09 (.10) [0.97]	.29 (.25) [2.70]	.12 (.16) [0.82]	.50 [0.97]
	3	.02 (.06) [0.26]	.10 (.14) [0.82]	.16 (.09) [1.00]	.28 [1.07]
		.20 [0.80]	.49 [0.98]	.32 [1.28]	

Note. *NEUS*: neuroticism self-rating; *NEUA*: neuroticism peer rating (A); log-linear parameters are presented in brackets. LEM requires dummy-coded latent two-variable parameters.

There is almost no method bias indicating that the latent marginal distributions do not differ from each other very strongly:

$$\begin{aligned} MB1_{(ns=1.na=1)} &= \frac{.23}{.20} = 1.15 \\ MB1_{(ns=2.na=2)} &= \frac{.50}{.49} = 1.02\,, \\ MB1_{(ns=3.na=3)} &= \frac{.28}{.32} = 0.88 \end{aligned} \qquad (5.1.19)$$

In addition to the method-bias the quasi-symmetry model also allows to examine the distinguishability of the latent categories (see Table 4.2.9). E.g.:

$$Dist_{(1.2)} = \frac{\hat{\pi}_{1.2}^{NEUS.NEUA}}{\hat{\pi}_{1}^{NEUS}\hat{\pi}_{2}^{NEUA}} = \frac{.10}{.49 \times .23} = 0.91, \qquad (5.1.20)$$

Table 5.1.10

Distinguishability indices for the two latent variables (NEUS and NEUA) in the quasi-symmetry latent rater agreement model for structurally different raters

		NEUA		
		1	2	3
	1		0.91	0.57
NEUS	2	0.90		0.75
	3	0.33	0.71	

Note. *NEUS*: neuroticism self-rating; *NEUA*: neuroticism peer rating (A)

The distinguishability indices show that self- and peer raters generally do not confound the categories of neuroticism (all indices are below 1). However, the distinguishability indices between the 1st and 2nd class (in either combination) are not very pronounced indicating that their joint expected proportions are almost as large as could be expected by chance. The distinguishability indices dealing with the 3rd class however show that this class is not confounded with any of the other two classes. This finding can be explained relying on the realistic accuracy model (Funder, 1995). Being traited (being neurotic) makes it much easier to be congruently (correctly) judged (see also Baumeister & Tice, 1988). Recall, that the latent one-variable parameters may differ and, therefore, the distinguishability indices also may differ.

$\kappa = .28$ indicates low agreement between the two raters. The ratios of expected cell proportions and the expected cell proportions based on the marginals are 1.8 for cell [1 1], 1.16 for cell [2 2], and 1.78 for cell [3 3]. That is, agreement is much higher for the 1st and 3rd class of neuroticism.

The latent one-variable model. The latent one-variable model fits to the data according to the bootstrapped χ^2-value. However, it fits worst with respect to the information criteria. The latent one-variable model will adequately represent the data if distinguishability and agreement are perfect (in this case the method bias type I will automatically be 1). Table 5.1.11 depicts the expected (conditional) proportions.

Table 5.1.11

Conditional response probabilities in the one-variable model for structurally different raters

		Latent status					
		n = 1		*n* = 2		*n* = 3	
variable		self	peer A	self	peer A	self	peer A
A / I (vulnerable)	1	.30	.27	.00*	.27	.03	.00*
	2	.42	.34	.10	.51	.04	.08
	3	.29	.39	.90	.22	.93	.92
B / J (sensitive)	1	.47	.42	.01	.38	.01	.00
	2	.45	.30	.01	.49	.10	.18
	3	.08	.28	.97	.14	.89	.82
C / K (moody)	1	.68	.63	.31	.64	.24	.39
	2	.21	.25	.31	.25	.28	.28
	3	.11	.12	.38	.10	.49	.32
D / L (doubtful)	1	.50	.61	.25	.63	.09	.21
	2	.28	.24	.12	.24	.17	.29
	3	.21	.15	.63	.12	.73	.50

Note. The column entitled self depicts the conditional response probabilities for the self-report; the column entitled peer A depicts the conditional response probabilities for the peer report.

Besides the worst information criteria, the one-variable model suffers from one major shortcoming in this application. The conditional response probabilities of peer *A* do not correspond to a typical response pattern for classes 1 and 2. The conditional response probabilities for items *C* and *D* are virtually identical and the conditional response probabilities for items *A* and *B* differ only to a small extent. Therefore, knowing only the peer ratings one could not differentiate between dyads (self- and peer raters) belonging to the first and second class. Therefore, the one-variable model does not represent the agreement structure in this application. This finding relates to the distinguishability indices found for the saturated latent rater agreement model. If a one-variable model fit the data the distinguishability indices should be very close to zero. This was by far not the case.

5.1.5 Comparison of the Latent Rater Agreement Models for Structurally Different Raters and Their Implications for the Analysis of Convergent and Discriminant Validity

All different latent rater agreement models fit to the empirical data according to the bootstrap procedure implemented in LEM. The BIC and AIC indices can be used to differentiate between them in terms of their parsimony and to choose the model with the best trade-off of absolute fit and parsimony. However, besides statistical analyses one should also take theoretical considerations into account to choose among the models. The latent saturated and the latent independence models may serve as two benchmarks representing the most flexible and most restrictive model at the latent level. All other models fall between these two models (except for the one-variable model).

Figure 5.2 shows the relation between the different models. All models are nested with respect to one common saturated model; therefore, one might want to apply a χ^2-difference test deciding which model fits best. However, for none of the models the empirical χ^2-value did follow its theoretical distribution but the values were on the edges of the parameter space (p = .00 or p = 1.00), in these cases χ^2-difference test does not work (see Dominicus, Skrondal, Gjessing, Pederson, & Palmgren, 2006).

All models can be used to determine the reliability of different indicators measuring one single categorical trait (see 4.1.1) and to analyze the agreement (convergence) between the two raters measuring one construct. Agreement can be determined calculating κ or the ratio of expected cell proportions to the expected cell proportions given the marginals. The overall agreement rates κ are very small for all models. The benchmarks for coefficient κ of manifest agreement tables may serve as a heuristic for the analysis of the latent joint distributions.

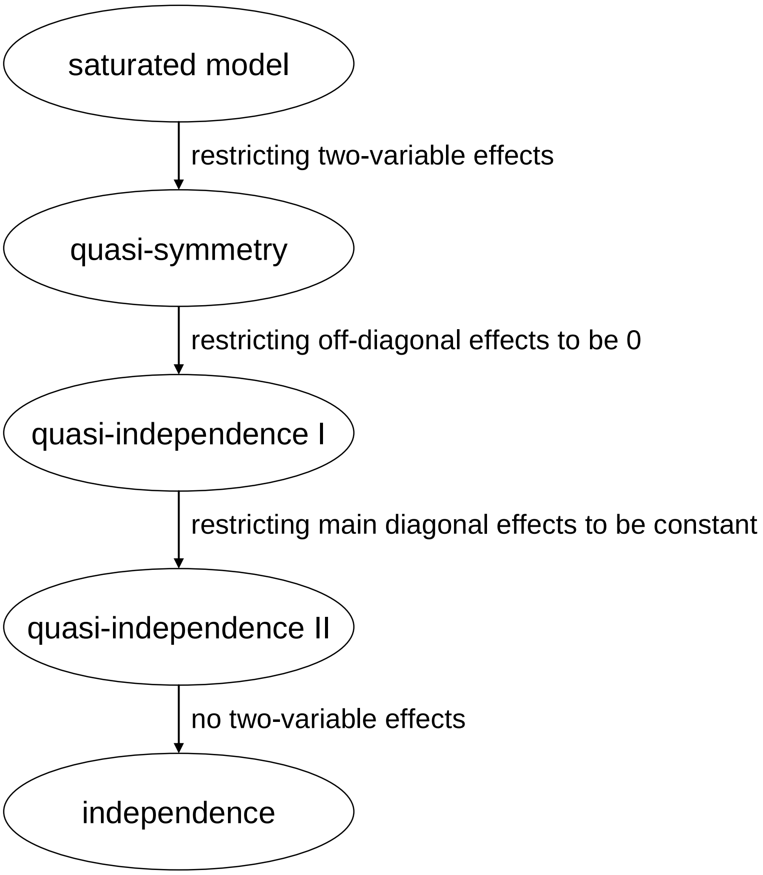
saturated model
restricting two-variable effects
quasi-symmetry
restricting off-diagonal effects to be 0
quasi-independence I
restricting main diagonal effects to be constant
quasi-independence II
no two-variable effects
independence

Figure 5.2. The relationship between the five latent rater agreement models presented in Section 5.1. Commentaries next to the arrows indicate the necessary constraints leading from one model to the other.

Another possibility is to inspect the latent two-variable log-linear parameters in models without boundary values. If there are boundary values the latent probability tables may be analyzed to get insight into the degree of agreement. Namely, the category-specific agreement ratios may be determined. The category-specific agreement ratios compare the product of the latent marginals to the model implied proportion for a particular cell. This corresponds conceptually to the calculation of χ^2-components in frequency tables (testing against independence). These values should be large to indicate high convergent validity. In the current applications the category-specific agreement rates roughly fall in the range of 1.2 and 1.9. Considering the relatively low expected proportions in the joint distributions, these values do not indicate large absolute agreement rates above the agreement expected by the product of the latent marginals. This is in line with the general finding that self- and peer raters do not agree to a large extent (see Funder, 1995).

The log-linear parameters reveal if agreement is constant across categories or if agreement is specifically high for some categories representing rather good categories of a trait. In the presented applications the quasi-independence II model fits best to the data implying that there is stable agreement across the latent categories between raters and that there are no specific patterns of disagreement.

Additionally, the models provide information about bias (as the difference between two presumed prevalence rates) and category distinguishability. Method bias can be determined calculating the *MB*1 coefficient. In the current applications, there is virtually no bias. To my mind no guidelines have been proposed concerning the amount of differences in latent prevalence rates to be considered severe. Zwick (1988) states that agreement should not be analyzed if the prevalence rates differ to a great extent. However, she does not give guidelines as to which difference one still may analyze rater agreement.

Category distinguishability may be calculated in the saturated and the quasi-symmetry model. These models imply, that self- and peer raters do not confound the 1st and 3rd latent categories of neuroticism. All other category combinations are expected to a lower degree than based on the marginals but the deviation is not very pronounced (the quasi-independence II structure is reflected). The distinguishability index shows if the two raters have tendencies to confound special categories with respect to the other rater's score. Distinguishability indices larger than 1 indicate a lack of convergent validity or a labeling problem. If there is agreement and, additionally, some categories besides the main diagonal are overrepresented the two raters have different conceptualizations of the construct. Special patterns of disagreement (high distinguishability indices) may reveal that two categories of a latent construct can be confounded more easily than other categories. This may be due to an imperfect description of the categories but also be due to related yet still distinct categories (e.g., a gambling personality type may be confounded with a risk seeking personality type but probably not with a security oriented personality type). If these related categories are part of the latent cross-classification it may occur that there are systematic patterns of disagreement. The combination of gambling and risk-seeking may occur more frequently than expected by chance, whereas the two categories are rarely (less frequently than predicted by chance) confounded with the security oriented personality type. In the quasi-symmetry model, these overrepresentations are constant irrespective of the ordering of the raters (the effects are identical for the combination of "gambling and risk-seeking" as well as "risk-seeking and gambling"). If there are only high distinguishability indices but no agreement, it is very probable that a labeling problem occurred and one may check if the latent categories are ordered in the same way for the two raters.

The quasi-independence I model imposes relatively strong constraints on the model parameters. The only fitted cell frequencies in models for observed data are those of the main diagonal. Transferring this model to the latent level bears the difficulty to clearly interpret the log-linear parameters. There is no very clear substantive interpretation. However, the latent proportions can be easily interpreted. Comparing the expected cell proportions to the expected values given the marginals only gives a lower boundary for the reliability estimate of Schuster and Smith (2006). Additionally, this comparison shows the amount of agreement between raters. One may also consider κ for analyzing agreement at the latent level. Additionally, the ratio of expected proportions to the product of the latent marginals reveals the degree to which these categories are overrepresented. If the parameters for agreement on the main diagonal differ vastly from each other, agreement is category specific. That is, raters agree with each other also as a function of the category. It may be that some types (e.g., not neurotic) may be more easily identified than others and that, therefore, raters agree more often with respect to this category than with respect to other categories.

Cells besides the main diagonal must be underrepresented with respect to the product of their latent marginals. By model definition these cells do not show problems related to distinguishability or confounding of categories since the quasi-independence I model assumes disagreement to follow the assumption of independence. If rater agreement is not category specific but constant across all cells on the main diagonal the quasi-independence II model will fit to the data (this is the case in the current application). In this model, rater agreement is a property of the pair of raters. In both quasi-independence models rater bias (as difference in the prevalence rates) can be analyzed. Comparing the quasi-independence I and the quasi-independence II models reveals if the moderators good judge and good category (trait) interact (see Funder, 1995). They do so in the quasi-independence I model they do not in the quasi-independence II model.

The most restricted model is the independence model. In this model, there is no relationship between the two raters, that is, the only agreement between two raters is due to chance agreement. The raters do not have the slightest view in common with respect to the target. In general, this model will not fit to the data but may be analyzed to provide a lower boundary for the cross-classification of the latent joint distribution.

All models (but the independence model) share that agreement between raters can be modeled. Agreement is high if the log-linear two-variable effect(s) on the main diagonal are large. In general, the two-variable effects besides the main diagonal are expected to be smaller than 1 (their expected proportions should be smaller than the product of their marginal proportions). If there are large effects besides the main diagonal this may point to two different situations:

1. The patterns of the conditional response probabilities are similar across raters suggesting that the labels of the latent categories have well been chosen. In this case, one rater perceives completely different "behavioral cues" to judge the target person than the other. An investigation of the decision making process (e.g., Wickens, 2002) and determinants as well as moderators of agreement (Funder, 1995) might give more insight into these issues.
2. It turns out that the labels of the latent categories have not well been chosen. This may be due to relatively low reliabilities of the indicators which do not permit to clearly label the latent categories. The interpretation of the model must be carried out very carefully. If the latent categories are reliably measured it may be the case that either the latent categories are related in an unexpected way indicating very low convergent validity or a simple labeling problem occurred. Reconsidering the ordering of the classes might remedy the problem.

5.2 Latent Rater Agreement Models for Interchangeable Raters

Analyzing the convergence (agreement) of interchangeable raters for multiple items can also be done adopting the existing rater-agreement models to the latent level. Since interchangeable raters originate in the same distribution, the model parameters must be identical across raters. This implies measurement invariance (see below), identical prevalence rates, and, additionally, identical log-linear parameters for interchanged categories $\left(\tau_{x.y}^{X.Y} = \tau_{y.x}^{X.Y}\right)$. The two-variable effect describing the interaction of the latent categories not neurotic rated by *A* with moody but stable rated by *B*, for example, is identical to the inversed interaction not neurotic by *B* and moody but stable for *A* (therefore *x* and *y* are inversed on the right hand side of $\left(\tau_{x.y}^{X.Y} = \tau_{y.x}^{X.Y}\right)$.

Figure 5.3 presents a categorical monotrait-multimethod model for the analysis of latent rater agreement of two raters. For sake of comprehensibility the trait variables and the items are labeled. *NEUA* and *NEUB* represent the latent construct (*NEUA* for rater *A* vs. *NEUB* for rater *B*); the latent traits (class variables) are measured by the same set of items ("vulnerable, sensitive, moody, and self-doubtful").

I consider a total of three different manifest rater agreement models which can be adopted and defined for the analysis of latent rater agreement: In 5.2.2, the latent quasi-independence I rater agreement model, in 5.2.3, the latent quasi-independence II rater agreement model, and in 5.2.4, the latent symmetry rater agreement model will be defined.

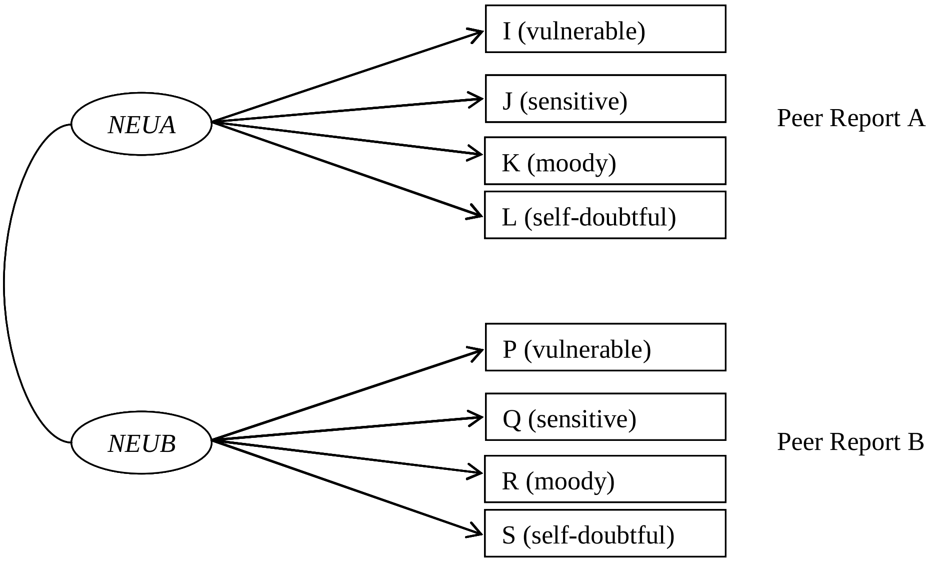

Figure 5.3. Log-linear model with two latent variables representing the latent construct Neuroticism (*NEU*) for the two peer reports *A* and *B*. Each latent variable is measured by four manifest indicators.

5.2.1 Measurement Invariance for Interchangeable Raters

Measurement invariance ensures that the link-function describing the genesis of the latent variables as representations of the joint observed ratings is the same for the two methods.

Definition 5.2.1

Measurement invariance for interchangeable raters

Let raters *A* and *B* be interchangeable due to theoretical reasons. Their latent variables *X* and *Y* representing the classification of *A*'s and *B*'s ratings (of the same construct; e.g., neuroticism: *NEU*) must fulfill the following restrictions:

i) identical number of latent categories

$$\max(x) = \max(y) = C\,. \qquad (5.2.1)$$

The maximum number of categories is the same for the two ratings.

ii) identical latent distributions

$$\tau_x^X = \tau_y^Y\text{, for } x = y \qquad (5.2.2)$$

for latent categories representing the identical latent category.

iii) identical link functions

$$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X} = \mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y} \text{ with } \tau_{m_i}^{M_i} = \tau_{o_i}^{O_i} \wedge \tau_{m_i.x}^{M_i.X} = \tau_{o_i.y}^{O_i.Y}\,, \qquad (5.2.3)$$

for m_i and o_i representing identical categories of identical items and $x = y$.

Explanation:

i) The number of categories must be the same for the two latent variables (*X* and *Y*) because the two originate in the same population.

ii) Therefore their rating also show identical prevalence rates (see also Schuster & Smith, 2002, 2006; Zwick, 1988).

iii) Identical link-functions produce identical expected manifest response patterns given identical latent statuses for the two raters (see Eid, Langeheine, & Diener, 2003 for a related topic in cross-cultural psychology) which must also be the case due to the interchangeability (the random sampling out of one set).

Measurement invariance ensures, that the links of manifest indicators to the latent variables are the same, and the observed responses follow the same distributions. Measurement invariance does not imply that different raters (methods) provide the same scores / ratings given a particular target. Identical ratings can only be observed in the case of perfect agreement between raters.

5.2.2 Definition of the Quasi-Independence I Latent Rater Agreement Model for Interchangeable Raters

Definition 5.2.2

The latent quasi-independence I model for interchangeable raters

Let *X* and *Y* represent the same latent construct measured by two interchangeable raters.

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \left(\tau_{x.y}^{X.Y}\right)^I, \quad \text{with} \begin{cases} I = 1 \text{ if } x = y \\ I = 0 \text{ else} \end{cases} \qquad (5.2.4)$$

with $\mathrm{T}_{\mathbf{a}} = \mathrm{T}_{\mathbf{b}}$ following Equation 5.2.3, $\tau_x^X = \tau_y^Y$ for $x = y$, and $\tau_{x.y}^{X.Y} = \tau_{y.x}^{X.Y}$ for $x = y$.

$e_{\mathbf{ab}.ns.cs}$ is the expected frequency of a specific cell in the latent joint cross-classification of the manifest response patterns $\mathbf{ab}$ (consisting of the two rater-specific patterns $\mathbf{a}$ and $\mathbf{b}$) with the two latent variables *X* and *Y*. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the product of the log-linear parameters linking the latent variable *X* to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the product of the log-linear parameters linking the latent variable *Y* to its indicators and the manifest one-variable effects.

$\mathrm{T}_{\mathbf{a}} = \mathrm{T}_{\mathbf{b}}$ implies that identically worded items have identical model parameters.

τ_x^X and τ_y^Y: represent the latent one-variable parameters. $\tau_{x.y}^{X.Y}$: represent the latent two-variable parameters.

The statistical meaning of the model parameters and their implications are identical to the meaning of the model parameters of the latent quasi-independence I model for structurally different raters.

5.2.3 Definition of the Quasi-Independence II Latent Rater Agreement Model for Interchangeable Raters

Definition 5.2.3

The latent quasi-independence II model for interchangeable raters

Let *X* and *Y* represent the same latent construct measured by two interchangeable raters.

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \left(\tau_{x.y}^{X.Y}\right)^I, \quad \text{with } \begin{cases} I = 1 \text{ if } x = y \\ I = 0 \text{ else} \end{cases}, \qquad (5.2.4 \text{ repeated})$$

and $\tau_{x.y}^{X.Y} = \tau^{X.Y}$.

with $\mathrm{T}_{\mathbf{a}} = \mathrm{T}_{\mathbf{b}}$ following Equation 5.2.3, $\tau_x^X = \tau_y^Y$ for $x = y$, and $\tau_{x.y}^{X.Y} = \tau_{y.x}^{X.Y}$ for $x = y$.

$e_{\mathbf{ab}.ns.cs}$ is the expected frequency of a specific cell in the latent joint cross-classification of the manifest response patterns **ab** (consisting of the two rater-specific patterns **a** and **b**) with the two latent variables *X* and *Y*. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the product of the log-linear parameters linking the latent variable *X* to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the product of the log-linear parameters linking the latent variable *Y* to its indicators and the manifest one-variable effects.

$\mathrm{T}_{\mathbf{a}} = \mathrm{T}_{\mathbf{b}}$ implies that identically worded items have identical model parameters.

τ_x^X and τ_y^Y: represent the latent one-variable parameters. $\tau_{x.y}^{X.Y}$: represent the latent two-variable parameters.

The statistical meaning of the model parameters and their implications are identical to the meaning of the model parameters of the latent quasi-independence II model for structurally different raters.

5.2.4 Definition of the Symmetry (Saturated) Latent Rater Agreement Model for Interchangeable Raters

Definition 5.2.4

The latent symmetry model for interchangeable raters

Let *X* and *Y* represent the same latent construct measured by two interchangeable raters.

$$e_{\mathbf{ab}.x.y} = \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \tau_x^X \tau_y^Y \tau_{x.y}^{X.Y}, \qquad (5.2.5)$$

with $\mathrm{T}_{\mathbf{a}} = \mathrm{T}_{\mathbf{b}}$ following Equation 5.2.3, $\tau_x^X = \tau_y^Y$ for $x = y$, and $\tau_{x.y}^{X.Y} = \tau_{y.x}^{X.Y}$ for $x = y$.

$e_{\mathbf{ab}.ns.cs}$ is the expected frequency of a specific cell in the latent joint cross-classification of the manifest response patterns $\mathbf{ab}$ (consisting of the two rater-specific patterns $\mathbf{a}$ and $\mathbf{b}$) with the two latent variables *X* and *Y*. η is the overall geometric mean of the complete table (manifest and latent variables). $\mathrm{T}_{\mathbf{a}}$ and $\mathrm{T}_{\mathbf{b}}$ represent the measurement models of the latent variables:

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.x}^{M_i.X}$: represents the product of the log-linear parameters linking the latent variable *X* to its indicators and the manifest one-variable effects,

$\mathrm{T}_{\mathbf{b}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.y}^{O_k.Y}$: represents the product of the log-linear parameters linking the latent variable *Y* to its indicators and the manifest one-variable effects.

$\mathrm{T}_{\mathbf{a}} = \mathrm{T}_{\mathbf{b}}$ implies that identically worded items have identical model parameters.

τ_x^X and τ_y^Y : represent the latent one-variable parameters.

$\tau_{x.y}^{X.Y}$: represent the latent two-variable parameters.

In the case of interchangeable raters the symmetry and saturated model are identical. Since measurement invariance must hold and additionally the latent one-variable effects are restricted to be equal across raters the latent marginals must be identical. Moreover, since the two raters are interchangeable their disagreement must follow the assumption of (quasi-) symmetry.

5.2.4.1 Implications of the symmetry model for interchangeable raters

The latent one-variable effects reflect the univariate latent distributions of the latent variables. There are no differences between the latent distributions by definition. Therefore, none of the ratings is biased with respect to the other rating. However, the ratings can be biased with respect to the true prevalence rates of the construct.

In the symmetry model, the distinguishability index can be estimated as in the quasi-symmetry model. The ratio to which proportions of specific cell-combinations besides the main diagonal deviate from the expected proportions given the one-variable effects is defined as distinguishability index.

Definition 5.2.5

Distinguishability index (Dist) for interchangeable raters

$$Dist_{(x.y)} = \left(\tau^{X.Y}_{x.y}\right) = \left(\tau^{X.Y}_{y.x}\right) = \sqrt{\frac{\pi^{X.Y}_{x.y}\pi^{X.Y}_{y.x}}{\pi^{X}_{x}\pi^{Y}_{y}\pi^{X}_{y}\pi^{Y}_{x}}} = \frac{\pi^{X.Y}_{x.y}}{\pi^{X}_{x}\pi^{Y}_{y}} \text{, for } x \neq y . \tag{5.2.6}$$

Lemma for Definition 5.2.5:

$$\begin{aligned} &\pi^{X}_{x} = \pi^{Y}_{x} \text{ since } \tau^{X}_{x} = \tau^{Y}_{x} \\ &\pi^{Y}_{y} = \pi^{X}_{y} \text{ since } \tau^{X}_{y} = \tau^{Y}_{y} \qquad \text{, (by definition 5.2.4)} \\ &\pi^{X.Y}_{x.y} = \pi^{X.Y}_{y.x} \text{ since } \tau^{X.Y}_{x.y} = \tau^{X.Y}_{y.x} \end{aligned} \tag{5.2.7}$$

Replacing:

$$\sqrt{\frac{\pi^{X.Y}_{x.y}\pi^{X.Y}_{y.x}}{\pi^{X}_{x}\pi^{Y}_{y}\pi^{X}_{y}\pi^{Y}_{x}}} = \sqrt{\frac{\left(\pi^{X.Y}_{x.y}\right)^2}{\left(\pi^{X}_{x}\right)^2\left(\pi^{Y}_{y}\right)^2}} = \frac{\pi^{X.Y}_{x.y}}{\pi^{X}_{x}\pi^{Y}_{y}} \tag{5.2.8}$$

The distinguishability index for interchangeable raters shows to which ratio particular cells of the joint distribution representing discordant ratings are over- or underrepresented. Due to the interchangeability of raters, this coefficient must yield identical results for cells mirrored at the main diagonal. Values larger than one indicate that the proportions of the cell combinations *x.y* and *y.x* are higher than expected by chance, values smaller than one indicate that these proportions are smaller than expected by chance. If the values are larger than one, the two raters confound the categories *x* and *y*. That is, if one of them chooses category *x* the probability to observe category *y* for the other rater increases. If the index is smaller than one the two raters produce smaller latent proportions for these cells than expected on chance and, therefore, one may conclude that they distinguish well between these categories. Distinguishability indices larger than 1 indicate a lack of discriminant validity. The distinguishability index may be related to moderators of agreement (accurate) ratings as described for the quasi-symmetry model for structurally different raters. Focusing on disagreement, researchers might use this information to inspect if possible moderators influence the high or low disagreement rates.

Focusing on agreement, the category-specific agreement rates may be used to show for which categories high agreement could be obtained.

Congruent ratings which go beyond the agreement by chance are reflected by the two-variable effects on the main diagonal $\left(\tau_{x.y}^{X,Y}\right)$ for $x = y$. These effects show agreement between raters. Agreement for raters is a special case of convergence in general. Since these parameters may differ between cells on the main diagonal these parameters depict the category-specific convergence beyond chance convergence. Additionally κ and the category-specific agreement ratios can be calculated.

5.2.5 Applications of the Latent Rater Agreement Models for Interchangeable Raters

The latent rater agreement models for interchangeable raters will be illustrated relying on the empirical example of neuroticism measured by the two peer-reports *A* and *B*. The data have been described in Section 4.1. The two raters use exactly the same items (“vulnerable”, “sensitive”, “moody”, and “self-doubtful”) and response categories (low, middle, high). Moreover, the peer raters have been randomly assigned to be peer *A* and peer *B* yielding interchangeable raters. Measurement invariance must thus hold across raters.

Table 5.2.1

Goodness-of-fit coefficients of the different latent rater agreement models for interchangeable raters

									Bootstrap
Model	χ^2	$p\left(\chi^2\right)$	L^2	p(L^2)	Df	AIC^1	BIC^1	boundaries	p_{bootP}
Saturated (symmetry) model	6492.45	.63	1620.18	1.00	6531	–11441.82	–38632.43	2	.40
Independence model	6659.81	.14	1650.54	1.00	6534	–11417.46	–38620.56	2	.36
Quasi-independence I model	6627.78	.20	1631.83	1.00	6532	–11432.17	–38626.95	—	.45
Quasi-independence II model	6471.99	.70	1623.64	1.00	6533	–11442.36	–38641.30	—	.39
One-variable model	7677.76	.00	1805.65	1.00	6534	–11262.35	–38465.45	—	.10

Note: χ^2: Pearson χ^2-value; L^2 likelihood-based χ^2-value; ^{1}AIC and BIC are based on L^2-values; boundaries: number of boundary values; the bootstrap consisted of 200 bootstrap samples, p_{bootP}: bootstrapped Pearson χ^2-value.

Table 5.2.1 presents the goodness-of-fit coefficients for the different latent rater agreement models for interchangeable raters. The empirical χ^2-values do not follow their theoretically expected distributions, therefore, the bootstrapped *p*-values should be examined. According to the bootstrap all models fit to the data. The quasi-independence II model fits best to the data according to the AIC and BIC criteria. Moreover, this model does not suffer from any boundary value. Since the models for interchangeable raters do not differ in their interpretation from the models for interchangeable raters only the quasi-independence II model will be discussed.

Table 5.2.2

Log-linear parameters of the measurement model of the latent quasi-independence-II latent rater agreement model

		one-variable effect	two-variable effect[1]		
variable	manifest categories		*na* = *nb* = 1	*na* = *nb* = 2	*na* = *nb* = 3
I / *P* (vulnerable)	1	0.396	8.958	0.146	0.766
	2	1.311	1.488	3.196	0.210
	3	1.928	0.075	2.146	6.212
J / *Q* (sensitive)	1	0.659	4.439	0.923	0.244
	2	1.232	1.272	1.477	0.532
	3	1.231	0.177	0.733	7.700
K /*R* (moody)	1	2.073	1.490	1.301	0.516
	2	0.890	0.914	1.167	0.938
	3	0.542	0.735	0.659	2.068
L / *S* (doubtful)	1	1.572	2.146	0.981	0.475
	2	0.871	0.833	1.225	0.980
	3	0.731	0.560	0.832	2.147

Note. [1] *na*: latent category of *NEUA*; *nb*: latent category of *NEUB*. *I* through *L*: Items measuring *NEUA*; *P* through *S*: Items measuring *NEUB*.

Table 5.2.2 presents the log-linear parameters of the quasi-independence II model. These parameters are identical for peer reports *A* and *B* since the two peer reports are interchangeable. The pattern of log-linear parameters fits well to the results of the previously reported results for models with interchangeable raters. The 1st latent class of variable *NEUA* or *NEUB* representing the target's latent neuroticism score rated by peer *A* or peer *B* is characterized by high two-variable log-linear parameters for the 1st manifest response category. The log-linear parameters linking the 2nd manifest response category to the 1st latent class are also larger than 1 in two cases (for items “vulnerable” and “sensible”). Raters belonging to this class thus generally prefer the first manifest response category compared to their response tendencies for the other latent classes. In order to determine if they absolutely prefer the 1st response category the manifest one-variable parameters must also be considered. This is done in calculating the conditional response categories presented in Table 5.2.3.

The 2nd latent class is characterized by large two-variable effects linking the 2nd manifest category to the latent class for items vulnerable and sensitive. For item vulnerable also very large effects can be found for the 3rd manifest response category. The two-variable parameters for moody and self-doubtful do not vary much across their manifest categories.

The 3rd latent class shows very large two-variable log-linear parameters for the 3rd manifest category. These values are always higher than those of the other two classes. Moreover, the two-variable parameters for the 1st and 2nd manifest category are always smallest for the 3rd latent category compared to the 1st and 2nd latent category. Table 5.2.3 presents the conditional response probabilities. Inspecting the conditional response probabilities reveals the same results as in the models combining self- and peer report *A*.

Table 5.2.3

Conditional response probabilities of the manifest response categories in the latent quasi-independence-II latent rater agreement model

		latent status		
variable	manifest categories	*na* = *nb* = 1	*na* = *nb* = 2	*na* = *nb* = 3
I / P (vulnerable)	1	.63	.01	.02
	2	.35	.50	.02
	3	.03	.49	.95
J / Q (sensitive)	1	.62	.18	.02
	2	.33	.55	.06
	3	.05	.27	.92
K / R (moody)	1	.72	.66	.35
	2	.19	.25	.28
	3	.09	.09	.37
L / S (doubtful)	1	.75	.48	.24
	2	.16	.33	.27
	3	.09	.19	.50

Note. *na*: latent category of *NEUA*; *nb*: latent category of *NEUB*. *E* through *H*: Items measuring *NEUA*; *I* through *L*: Items measuring *NEUB*.

The latent quasi-independence II model implies that the overrepresentation of the agreement cells on the main diagonal is constant. In this application, agreement is (constantly) 2.17 times more frequent than expected based on the product of the latent one-variable parameters.

Table 5.2.4

Latent joint distribution of the quasi-independence II latent rater agreement model

	nb = 1	*nb* = 2	*nb* = 3	
na = 1	.10 (.06) [2.17]	.07 (.10)	.07 (.09)	.24 [2.17]
na = 2	.07 (.10)	.23 (.16) [2.17]	.10 (.15)	.40 [1.07]
na = 3	.07 (.09)	.10 (.15)	.20 (.14) [2.17]	.37 [1.00]
	.24 [0.71]	.40 [1.07]	.37 [1.00]	1

Note. *na*: latent category of *NEUA*; *nb*: latent category of *NEUB*. Values in parentheses represent the expected values given the latent marginals only. Values in brackets represent the dummy coded log-linear parameters.

Table 5.2.4 presents the latent joint distribution of the quasi-independence II latent rater agreement model. As can be seen, the latent joint distribution is mirrored around the main diagonal as a particular consequence of equal latent marginal distributions (this is also true for the independence and the quasi-independence I models). Although the agreement rate is modeled using a constant parameter, this does not imply that ratio of the expected proportions of cells on the main diagonal to their expectancies given the marginals is constant. In fact these ratios are 1.67 for [1 1], 1.44 for [2 2], and 1.43 for [3 3]. However, the ratios differ to a smaller extent than for the quasi-independence I model. This is an effect of the constant two-variable parameter. $\kappa = .27$ indicates poor rater agreement.

LEM does only allow for a specification of dummy coded log-linear parameters. Therefore, the parameters cannot be interpreted as in the model definition. They can be interpreted as deviations from the reference category for ambiguous cases and as indicators of the constant latent class size for obvious cases.

5.2.6 Implications of the Rater Agreement Models for Interchangeable Raters

In this section, latent rater agreement models have been defined for the analysis of one construct measured by two interchangeable raters. As for structurally different raters, all models are nested with respect to one model—the latent symmetry model (see Figure 5.4). The empirical χ^2-values did not follow their theoretical distributions, therefore, I do not compute the χ^2-difference tests (Dominicus et al., 2006).

The convergent validity of two or more methods (raters) in measuring the same trait can also be examined using overall agreement indices as κ (in models allowing for higher rates of agreement). κ indicated rather poor agreement rates with respect to the benchmarks for manifest agreement.

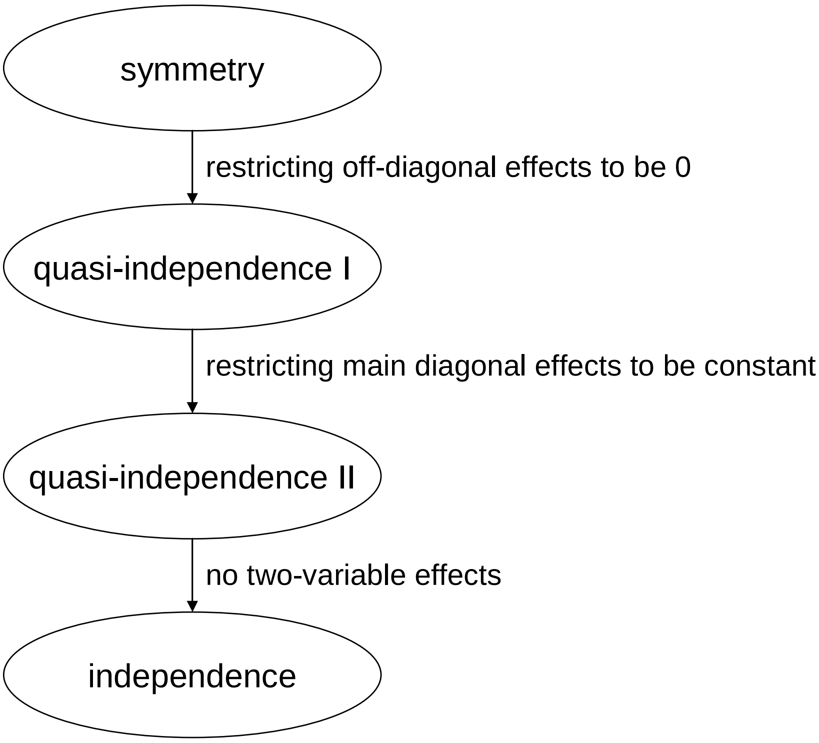
symmetry
restricting off-diagonal effects to be 0
quasi-independence I
restricting main diagonal effects to be constant
quasi-independence II
no two-variable effects
independence

Figure 5.4. The relationship between the four latent rater agreement models for interchangeable raters (except for the one-variable model) presented in Section 5.2. Commentaries next to the arrows indicate the necessary constraints leading from one model to the other.

In the current applications, the category-specific agreement rates fell into the range of 1.4 to 1.7 indicating relatively low agreement on neuroticism given the low products of the latent marginals. These values reveal if agreement is constant across categories.

All models presented in this section fit to the empirical data indicating their applicability. However, as could be shown calculating coefficient kappa or by an inspection of the latent proportions agreement is not very pronounced and disagreement does not differ from chance disagreement (quasi-independence assumption).

5.3 Discussion of the Latent Rater Agreement Models for Structurally Different and Interchangeable Raters

In this chapter, manifest rater agreement models have been adapted to the level of latent variables. These models allow examining latent typologies, that is, agreement between raters can be determined for more than one observed variable per TMU. It is the response pattern that determines the membership to a latent class, agreement is no longer bound to the more error prone single classification on single items.

Moreover, the models allow for reducing complex sets of rater agreement data. Imagine, a complete data set of two raters using two times four items to rate two clinical disorders. Comparing the data at the observed level would result in a comparison of 4 x 4 = 16 agreement tables. The models presented here allow reducing the information to be compared to a finite (and usually small) number of classes. If the model-implied typology corresponds to the data and the mean assignment probabilities are rather high (or the strength of the relation between latent and manifest variables is high) it is useful, parsimonious, and efficient to consider agreement at the latent level. In empirical applications, a cross-validation of the results found for the latent rater agreement models by estimations of other models is needed to guarantee that the model results are correct.

In principle, the models allow for a test or for the explorative analysis if raters are interchangeable or not (restricting their measurement models). Additionally, one can analyze if the raters confound particular categories or if they can well distinguish between all categories. This analysis can be carried out comparing different models which imply different patterns of agreement and disagreement but also by an inspection of the distinguishability index. The distinguishability index is newly introduced. The fact that raters confound particular categories may be of interest in training programs for clinical psychologists, for example, in order to achieve a fine-graded distinction between clinical symptoms (as latent classes), and this may also be of interest in research programs on rater accuracy (Funder, 1995).

The models also principally allow for an inspection of determinants or moderators of agreement and disagreement (see Funder, 1995). Focusing on disagreement, researchers might use the distinguishability index to inspect which disagreement cells are overrepresented. Incorporating additional variables into the model may help to explain this effect (see Section 6 for an additional construct). Focusing on agreement, the category-specific agreement rates may be used to show for which categories high agreement could be obtained yielding some information about the moderator good trait or about which category of a trait is a good category.

In order to additionally analyze the discriminant validity of different latent typologies and to shed some light on personality traits that could enhance agreement on other traits, the latent rater agreement models for one construct and two methods (Monotrait-Multimethod models) have to be extended to the analysis of more than one construct. The next section defines and illustrates the resulting Multitrait-Multirater models.

6 Correlated Traits Multitrait-Multirater Model

In this chapter the previously described saturated and symmetry models for two latent variables will be extended to the simultaneous analysis of 2 traits and 2 raters yielding the Correlated Traits Multitrait-Multirater (CT MTMR) model. This model allows for analyzing structurally different as well as interchangeable raters. The model will first be defined for the case of structurally different raters. The model for interchangeable raters emerges imposing the measurement invariance and necessary interchangeability restrictions (see Section 5.2). The saturated log-linear model with four latent variables will be formally defined and its parameters will be related to the criteria of convergent and discriminant validity presented by Campbell and Fiske (1959). I will indicate and introduce meaningful coefficients which indicate aspects of convergent and discriminant validity as well as aspects of method bias that are usually *not* addressed in MTMM analyses.

6.1 Definition of the Correlated Traits Multitrait-Multirater Model for Structurally Different Raters

In order to define the CT MTMR model the same prerequisites as described in Section 5.1 must be met. That is, all items belonging to the different trait-method-units (TMU) must be indicators of the constructs. Therefore, the two raters provide categorical ratings that can be categorized as described in Section 4.1 (separate log-linear models with one latent variable). The Monotrait-Multirater models allow for testing if the latent categories represent the same latent constructs and if the raters agree (convergent validity). The CT MTMR models allow for an additional analysis of discriminant validity. If the same construct is represented across raters this will result in similar latent categories across raters (see also Section 5.1) with similar meanings of the typical response patterns, similar relationships to other variables, and / or similar effects on other variables. The Correlated Traits Multitrait Multirater (CT MTMR) model for structurally different raters is a flexible model for the analysis of convergent and discriminant validity.

Definition 6.1.1

The saturated Correlated Traits Multitrait-Multirater (CT MTMR) model

Let *XS*, *XA* be two latent variables representing the same construct and let *YS*, *YA* be two other latent variables representing another construct. The latent variables representing different constructs are measured with distinct sets of items. *xs*, *xa*, *ys*, and *ya* indicate the latent categories of the four latent variables.

The saturated CT MTMR-model is defined as:

$$\begin{aligned} e_{\mathbf{abcd}.xs.xa.ys.ya} &= \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \mathrm{T}_{\mathbf{c}} \mathrm{T}_{\mathbf{d}} \tau_{xs}^{XS} \tau_{xa}^{XA} \tau_{ys}^{YS} \tau_{ya}^{YA} \\ &\times \tau_{xs.xa}^{XS.XA} \tau_{xs.ys}^{XS.YS} \tau_{xs.ya}^{XS.YA} \tau_{xa.ys}^{XA.YS} \tau_{xa.ya}^{XA.YA} \tau_{ys.ya}^{YS.YA} \\ &\times \tau_{xs.xa.ys}^{XS.XA.YS} \tau_{xs.xa.ya}^{XS.XA.YA} \tau_{xs.ys.ya}^{XS.YS.YA} \tau_{xa.ys.ya}^{XA.YS.YA} \tau_{xs.xa.ys.ya}^{XS.XA.YS.YA} \end{aligned} , \qquad (6.1.1)$$

with **abcd** being a transposed vector of observed responses, $\mathrm{T}_{\mathbf{a}}, \mathrm{T}_{\mathbf{b}}, \mathrm{T}_{\mathbf{c}}$, and $\mathrm{T}_{\mathbf{d}}$ representing the measurement models of the four latent variables (see also Equation 4.1.2):

η: is the geometric mean of the unobserved latent table (containing manifest and latent variables),

$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau_{m_i}^{M_i} \tau_{m_i.xs}^{M_i.XS}$: represents the log-linear parameters linking the latent variable *XS* to its indicators and the manifest one-variable effects, $\mathrm{T}_{\mathbf{b}} = \prod_{n_j=1}^{J} \tau_{n_j}^{N_j} \tau_{n_j.xa}^{N_j.XA}$: represents the log-linear parameters linking the latent variable *XA* to its indicators and the manifest one-variable effects, $\mathrm{T}_{\mathbf{c}} = \prod_{o_k=1}^{K} \tau_{o_k}^{O_k} \tau_{o_i.ys}^{O_k.YS}$: represents the log-linear parameters linking the latent variable *YS* to its indicators and the manifest one-variable effects, and $\mathrm{T}_{\mathbf{d}} = \prod_{p_l=1}^{L} \tau_{p_l}^{P_l} \tau_{p_l.ya}^{P_l.YA}$: represents the log-linear parameters linking the latent variable *YA* to its indicators and the manifest one-variable effects.

$\tau_{xs}^{XS}, \tau_{xa}^{XA}, \tau_{ys}^{YS}$, and τ_{ya}^{YA}: are the latent one-variable effects,

$\tau_{xs.xa}^{XS.XA}, \tau_{xs.ys}^{XS.YS}, \tau_{xs.ya}^{XS.YA}, \tau_{xa.ys}^{XA.YS}, \tau_{xa.ya}^{XA.YA}$, and $\tau_{ys.ya}^{YS.YA}$: are latent two-variable effects, $\tau_{xs.xa.ys}^{XS.XA.YS}, \tau_{xs.xa.ya}^{XS.XA.YA}$, : $\tau_{xs.ys.ya}^{XS.YS.YA}$, and $\tau_{xa.ys.ya}^{XA.YS.YA}$ are latent three-variable effects, $\tau_{xs.xa.ys.ya}^{XS.XA.YS.YA}$: represents latent four-variable effects.

6.1.1 The Statistical Meaning of the Different Effects in the Saturated CT MTMR Model

Figure 6.1 presents the CT MTMR model for two constructs measured by two structurally different raters (Definition 6.1.1). To make the presentation more comprehensible the latent variables are labeled representing neuroticism (*NEU*) and conscientiousness (*CON*) measured by a self-report (*S*) and a peer-report (*A*). The items correspond to the items of the empirical data described in Section 4.1. However, the model may also be estimated with more or fewer manifest variables.

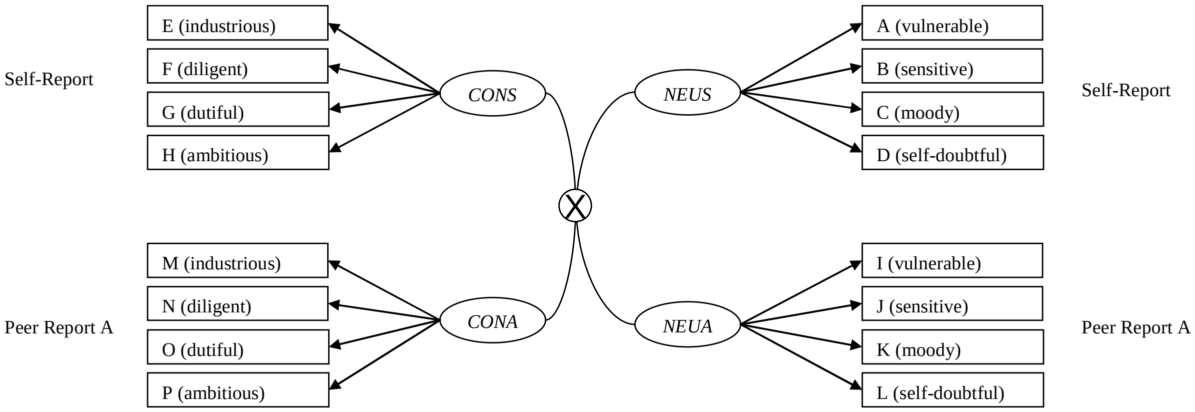

Figure 6.1 Categorical Multitrait-Multirater model for two traits measured by two raters. The ⊗ indicates hierarchical higher order effects (i.e., two-, three-, and four-variable effects).

The log-linear parameters of Definition 6.1.1 have the following meanings:

- η is the geometric mean of the unobserved complete frequency table.

- The submodels T_a, T_b, T_c, and T_d: have been described in section 4.1 (e.g., Goodman, 1974a, 1974b; Haberman, 1979; Hagenaars, 1990, 1993; McCutcheon, 1987).

- The latent one-variable parameters $\left(\tau^{XS}_{xs};\tau^{XA}_{xa};\tau^{YS}_{ys};\tau^{YA}_{ya}\right)$ represent the univariate distributions of the latent variables in the latent four-dimensional table. E.g.[14]:

$$\tau^{XS}_{xs}=\frac{\sqrt[YS\cdot YA\cdot XA]{\prod_{ys=1}^{YS}\prod_{ya=1}^{YA}\prod_{xa=1}^{XA}\pi^{XS.XA.YS.YA}_{xs.xa.ys.ya}}}{\sqrt[XS\cdot YS\cdot YA\cdot XA]{\prod_{w=1}^{XS}\prod_{ys=1}^{YS}\prod_{ya=1}^{YA}\prod_{xa=1}^{XA}\pi^{XS.XA.YS.YA}_{w.xa.ys.ya}}}, \qquad (6.1.2)$$

with *xs* indicating the particular latent category of *XS* and *w* indexing the first to the last category of *XS* in the denominator.

- The latent two-variable effects $\left(\tau^{XS.XA}_{xs.xa};\tau^{XS.YS}_{xs.ys};\tau^{XS.YA}_{xs.ya};\tau^{XA.YS}_{xa.ys};\tau^{XA.YA}_{xa.ya};\tau^{YS.YA}_{ys.ya}\right)$ indicate the deviations of particular cell proportions from the prediction based on the lower order effects. E.g.:

$$\tau^{XS.YS}_{xs.ys}=\frac{\sqrt[YA\cdot XA]{\prod_{ya=1}^{YA}\prod_{xa=1}^{XA}\pi^{XS.XA.YS.YA}_{xs.xa.ys.ya}}}{\eta*\tau^{XS}_{ys}\tau^{YS}_{ys}}, \qquad (6.1.3)$$

with $\eta*$ indicating the geometric mean of the latent table (the complete table can be collapsed across the manifest variables).

[14] *XS*, *XA*, *XB*, *YS*, *YA*, and *YB* represent latent variables but they also represent the highest category of the corresponding latent variable. However, this is only the case in connection with Greek symbols representing sums or products $(\Sigma \text{ or } \Pi)$

- The latent three-variable effects ($\tau^{XS.XA.YS}_{xs.xa.ys}; \tau^{XS.XA.YA}_{xs.xa.ya}; \tau^{XS.YS.YA}_{xs.ys.ya};$ and $\tau^{XA.YS.YA}_{xa.ys.ya}$) depict the deviations of particular cell proportions from the predictions based on all lower order effects in the different latent trivariate subtables. E.g.:

$$\tau^{XA.YS.YA}_{xa.ys.ya} = \frac{\sqrt[XS]{\prod_{w=1}^{XS} \pi^{XS.XA.YS.YA}_{w.xa.ys.ya}}}{\eta * \tau^{XA}_{xa} \tau^{YS}_{ys} \tau^{YA}_{ya} \tau^{XA.YS}_{xa.ys} \tau^{XA.YA}_{xa.ya} \tau^{YS.YA}_{ys.ya}}, \qquad (6.1.4)$$

with $\eta*$ indicating the geometric mean of the latent table and *xs*, *xa*, *ys*, and *ya* indicating the latent categories of *XS*, *XA*, *YS*, and *YA*. The one-and two-variable effects can be determined as described in Equations 6.1.2 and 6.1.3.

- The latent four-variable effect $\left(\tau^{XS.XA.YS.YA}_{xs.xa.ys.ya}\right)$ depicts the deviation of the expected proportion of a particular cell from the predictions based on all lower order effects in the complete (quadrivariate) table:

$$\begin{aligned}
\tau^{XS.XA.YS.YA}_{xs.xa.ys.ya} &= \frac{\pi^{XS.XA.YS.YA}_{xs.xa.ys.ya}}{\eta * \tau^{XS}_{xs} \tau^{XA}_{xa} \tau^{YS}_{ys} \tau^{YA}_{ya} \tau^{XS.XA}_{xs.xa} \tau^{XS.YS}_{xs.ys} \tau^{XS.YA}_{xs.ya} \tau^{XA.YS}_{xa.ys} \tau^{XA.YA}_{xa.ya} \tau^{YS.YA}_{ys.ya}} \\
&\times \frac{1}{\tau^{XS.XA.YS}_{xs.xa.ys} \tau^{XS.XA.YA}_{xs.xa.ya} \tau^{XS.YS.YA}_{xs.ys.ya} \tau^{XA.YS.YA}_{xa.ys.ya}} \qquad , \qquad (6.1.5) \\
&= \frac{\pi^{XS.XA.YS.YA}_{xs.xa.ys.ya}}{\pi^{XS.XA.YS}_{xs.xa.ys} \pi^{XS.XA.YA}_{xs.xa.ya} \pi^{XS.YS.YA}_{xs.ys.ya} \pi^{XA.YS.YA}_{xa.ys.ya}}
\end{aligned}$$

the lower order effects can be determined as described in Equations 6.1.2, 6.1.3, and 6.1.4.

6.1.1.1 The impact of the different log-linear effects on the analysis of convergent and discriminant validity

The saturated CT MTMR model is a flexible model for the analysis of convergent and discriminant validity of multiple ratings. Therefore, the inspection of convergent and discriminant validity does not only consist of the analysis of zero-order bivariate relationships but on the analysis of higher order effects. Additionally, the impact of different trait constellations on agreement and disagreement could principally be analyzed to inspect complex interactions of moderators of agreement. In a restricted version of the CT MTMR model with only two-variable effects, the associations between the latent categories can be analyzed on the bivariate level. This analysis comes close to an examination of the criteria developed by Campbell and Fiske (1959). The different log-linear parameters at the different levels of the interaction (two-, three-, and four-variable effects) may all have an impact on the convergent and discriminant validity. I will start by inspecting the impact of the highest order interaction passing to the lower order interactions. For sake of simplicity, I will exclude all higher order effects when I discuss the lower order effects in order to avoid a misinterpretation due to existing higher order effects.

Table 6.1.1

Extracted part of the latent joint distribution in the saturated CT MTMR model for three categorical latent variables with four-variable effects

					YA	
				1	2	3
xs =1	*xa* =1	*YS*	1	*A*	M	N
			2	O	*B*	*P*
			3	Q	R	*C*
	xa=2	*YS*	1	*G*	1	2
			2	3	*H*	4
			3	5	6	*I*
…	…	…	…	…	…	…
xs = 2	*xa* =1	*YS*	1	*J*	7	8
			2	9	*K*	10
			3	11	12	*L*
	xa=2	*YS*	1	*D*	S	T
			2	U	*E*	V
			3	W	Z	*F*
…	…	…	…	…	…	…

Note. Only the cell combinations for *XS* = 1, 2 *XA* = 1, 2 are depicted. The scheme applies to all other combinations of latent categories as well.

Table 6.1.1 depicts an extracted part of the latent joint distribution for latent variables with at least three categories. The cells of this table fall into three parts: a) Cells indicating agreement on both constructs (cells *A* through *F*; dark grey; i.e. *A* represents the category-combination [1 1 1 1]), b) cells indicating agreement on one construct (cells *G* through *L* for agreement on *YS* and *YA* (*G* represents the category-combination [1 2 1 1]) and cells *M* through *Z*[15] for agreement on *XS* and *XA* (*M* represents the category-combination [1 1 1 2]), light grey), and c) cells indicating disagreement on both of the constructs (numerated from 1 through 12).

All expected cell proportions are influenced by the complete set of one-, two-, three-, and four-variable effects. Saturated models do not impose restrictions on the log-linear parameters and, therefore, perfectly reproduce the frequency table. The latent log-linear parameters directly relate to the expected proportions of the latent table as shown in Equations 6.1.2 through 6.1.5 (see also Section 4.1.2). The four-variable log-linear effects have the following meaning with respect to the convergent and discriminant validity:

i) Four-variable effects

Complete agreement. The four-variable log-linear parameters representing agreement on *both* constructs (*A* through *F*) indicate the judgeability of the targets (Funder, 1995) with respect to the traits under consideration. If these effects are larger than 1 the corresponding expected cell proportions are higher than expected based on all lower order effects[16]:

$$\tau_{xs.xa.ys.ya}^{XS.XA.YS.YA} = \frac{\pi_{xs.xa.ys.ya}^{XS.XA.YS.YA}}{\pi_{xs.xa.ys}^{XS.XA.YS}\pi_{xs.xa.ya}^{XS.XA.YA}\pi_{xs.ys.ya}^{XS.YS.YA}\pi_{xa.ys.ya}^{XA.YS.YA}}, \tag{6.1.6}$$

for $xs = xa$ and $ys = ya$. Several constellations are possible:

- All four-variable parameters for complete agreement cells (*A* through *F*) are larger than 1 and of equal size $\tau_{xs.xa.ys.ya}^{XS.XA.YS.YA} = \tau_{(xs.xa.ys.ya)'}^{XS.XA.YS.YA}$ for all $(xs.xa.ys.ya) \neq (xs.xa.ys.ya)'$[17] with $xs = xa$ and $ys = ya$. This indicates that the convergence of the two raters is stable across the different category combinations. The odds to agree given the expected proportions based on lower order effects (see

[15] I left out *X* and *Y* numerating the cells to avoid confusion with the latent categories.

[16] I consider population parameters throughout this section.

[17] $(xs.xa.ys.ya)'$ indicates that at least one combination of $xs = xa$ or $ys = ya$ differs with respect to $(xs.xa.ys.ya)$.

Eq. 6.1.6) are identical on all category combinations indicating agreement on both constructs. This overall agreement rate may be due to two reasons (see Funder, 1995): There is a group of individuals who are easily judgeable (good targets) or the traits are especially visible in some targets (*palpability*). Since the agreement rate is constant the judgeability of the targets or the palpability does not depend on the scores on one of the latent constructs (it is constant across all categories).

- All four-variable parameters for complete agreement cells (*A* through *F*) are larger than 1 but differ from each other. In this case, the raters agree more often than expected based on the lower order effects. Judgeability of targets depends partly on their status on the latent variables. Individuals who belong to a special easily judgeable category of one trait can be more easily accurately (congruently) judged on a category of the other traits as well. In this case, judgeability (as palpability) is a property of different constellations of the latent categories.

  However, particular categories of the other trait may also serve as indicators of judgeability. A good example may be extraverted individuals who spend much time with their friends, overtly show their feelings, and comment on their thoughts. These individuals should be easily classifiable on other traits as agreeableness and neuroticism as well. Therefore, raters may have no difficulties classifying these individuals as extraverted and, additionally, on their different statuses of neuroticism and conscientiousness, for example. This effect may be weaker or stronger depending on the different categories. $\tau^{XS.XA.YS.YA}_{xs.xa.ys.ya} \leq \tau^{XS.XA.YS.YA}_{(xs.xa.ys.ya)'}$ for all $(xs.xa.ys.ya) \neq (xs.xa.ys.ya)'$ with $xs = xa$ and $ys = ya$. Being extraverted may be part of the properties characterizing good targets.

  If there are only few but very large four-variable parameters for complete agreement cells low discriminant validity on agreement ratings is found. The latent categories of the different construct partly overlap and cannot be considered very distinct from each other.

- All four-variable parameters for complete agreement cells (*A* through *F*) are larger than 1 and differ from each other as a function of categories of one trait. $\tau^{XS.XA.YS.YA}_{xs.xa.ys.ya} \leq \tau^{XS.XA.YS.YA}_{xs.xa.(ys.ya)'}$ for all $(xs.xa.ys.ya) \neq xs.xa.(ys.ya)'$ with

$xs = xa$ and $ys = ya$ or $\tau^{XS.XA.YS.YA}_{xs.xa.ys.ya} \leq \tau^{XS.XA.YS.YA}_{(xs.xa)'.ys.ya}$ for all $(xs.xa.ys.ya) \neq (xs.xa)'.ys.ya$ with $xs = xa$ and $ys = ya$. This effect is a special case of the previously described phenomenon. It may occur that for different levels of the target person's extraversion the raters have fewer problems to correctly (at least congruently) classify these targets on the other construct with the *same accuracy* for all categories of this other construct. In this case, extraversion can be regarded as an indicator of visibility / judgeability. Extraverted individuals may show visible cues of other traits and can, therefore, be easily judged on these traits as well.

- The four-variable parameters for complete agreement cells (*A* through *F*) may also be smaller than 1. For the corresponding cells agreement on both constructs is less frequently expected than predicted on the lower order effects. This result would be rather awkward but could be explained in cases when the latent cells indicate categories that are (partly) mutually exclusive in the raters view. This generally also indicates a lack of discriminant validity because these categories co-occur less frequently than expected. For example, in the analysis of the convergent and discriminant validity of ratings with respect to clarity of one's own feelings and expressivity of feelings (see Lischetzke & Eid, 2003 for a conceptualization of these constructs), the cell indicating agreement on "does not show feelings" and "is clear about feelings" can logically be underrepresented because somebody who does not show feelings cannot be judged to know about her or his feelings. In this case, this finding fits into theoretical considerations and is reasonable.

- The four-variable parameters for complete agreement cells (*A* through *F*) do not differ from 1. In this case the quadrivariate agreement can be explained by lower order effects of agreement (see discussion below).

Partial agreement. Four-variable parameters of cells indicating agreement on one construct but not on the other for the quadrivariate joint distribution (cells *J* through Z) represent a special kind of *rater bias*:

$$\tau^{XS.XA.YS.YA}_{xs.xa.ys.ya} = \frac{\pi^{XS.XA.YS.YA}_{xs.xa.ys.ya}}{\pi^{XS.XA.YS}_{xs.xa.ys}\,\pi^{XS.XA.YA}_{xs.xa.ya}\,\pi^{XS.YS.YA}_{xs.ys.ya}\,\pi^{XA.YS.YA}_{xa.ys.ya}}, \qquad (6.1.7)$$

for either $(xs = xa \text{ and } ys \neq ya)$ or $(xs \neq xa \text{ and } ys = ya)$. Different constellations may occur:

- The four-variable parameters of cells indicating agreement on one but not on the other trait (cells *J* through Z) are larger than 1. This finding can be interpreted in terms of rater bias. Although raters agree on one construct they disagree systematically on the other construct. This may be the case for raters who agree on a target person's extraversion but who have different views or theories about the relation between extraversion and intelligence, for example. One rater may assume that moderately extraverted individuals also tend to be more intelligent while the other assumes moderately extraverted individuals to be very intelligent. This effect is a four-variable effect if they use the same behavioral cues to identify the target's level of extraversion and relate this information to their judgment of intelligence. This kind of effect may account for all cells indicating partial agreement or only for particular cells.

- The four-variable parameters of (particular or all) cells indicating agreement on one but not on the other trait (cells *J* through Z) are smaller than 1. In this case, disagreement between the two raters with respect to specific category combinations is underrepresented if they agree on the other construct. This may be the case if agreement on one construct is very hard to achieve because the trait under consideration is not easily judgeable, if two raters agree on judging this difficult trait, they will most probably agree on more easy to judge traits as well and therefore the expected proportions of the disagreement cells for the latter construct are much smaller given agreement on the first trait. For example, it may be much more difficult to judge an individual's attitudes towards specific minorities (e.g., racist, neutral, positive, no opinion) than judging the same individual's extraversion. If raters agree on the presumably not overtly expressed attitude against minorities they will most probably also be able to judge the individuals score on an openly observable trait as extraversion.

  This effect thus shows (if there is agreement) that there is higher agreement on one construct (on all or on one category) if there is agreement on the other one.

The opposite does not necessarily have to be true. In this case, one construct (or specific cells of this construct) is more difficult to judge than (categories of) the other construct.

- All four-variable parameters of cells indicating agreement on one but not on the other trait (cells *M* through Z) do not differ from 1. In this case, agreement on one construct is not related to disagreement on the other construct.

Disagreement. The latent four-variable parameters of cells besides the agreement and partial agreement cells (1 to 12) represent influences which may be due to bias or to general disagreement:

$$\tau_{xs.xa.ys.ya}^{XS.XA.YS.YA} = \frac{\pi_{xs.xa.ys.ya}^{XS.XA.YS.YA}}{\pi_{xs.xa.ys}^{XS.XA.YS}\pi_{xs.xa.ya}^{XS.XA.YA}\pi_{xs.ys.ya}^{XS.YS.YA}\pi_{xa.ys.ya}^{XA.YS.YA}}, \tag{6.1.8}$$

for $\left(xs \neq xa \text{ and} \, ys \neq ya\right)$. The following different constellations are possible:

- All four-variable parameters for complete disagreement cells (1 to 12) are larger than 1. In this case the two raters disagree more often than predicted based on the lower order effects. In general, this indicates a lack of convergent validity. κ for the quadrivariate joint distribution will be negative (κ may be determined as depicted in Section 2 considering only cells representing complete agreement (e.g., [1 1 1 1]) and their latent univariate marginals). However, there still might be a few positive category-specific agreement ratios for some cells. I do not expect this constellation to appear in any application. This constellation may appear in cases where raters do not follow their instructions or due to a wrong labeling of categories. Even if raters are guessing they should have four-variable parameters that do not differ from 1.

- Some (one) four-variable parameters for complete disagreement cells (1 to 12) are larger than 1. In this case particular combinations of one rater's latent scores are associated to the other rater's scores but for different cell combinations. If raters weigh some behavioral cues in different ways given cues on the other trait they may be more often categorized in latent disagreement cells. If, for example, one rater classifies an individual due to specific behavioral cues as highly extraverted and, additionally, these cues may lead this rater to also classify this individual as moderately neurotic this combination of behavioral cues may be associated to the moderately extraverted and highly neurotic classes for the other rater.

- Some (all) four-variable parameters for complete disagreement cells (1 to 12) are smaller than 1. This may in most cases be due to higher complete and / or partial agreement rates because the log-linear parameters are effect coded. Therefore, higher agreement also affects the disagreement cells in the saturated model. Yet, this may also be due to high disagreement on a particular cell combination and no effects on complete or partial agreement cells.

- None of the four-variable parameters for complete disagreement cells (1 to 12) differs from 1. In this case, there is neither an over- nor an underrepresentation of complete disagreement cells.

At the level of four-variable effects, there are some combinations of the above mentioned constellations that merit special attention because these can be related to the concepts of convergent and discriminant validity.

Overall agreement may be high due to bivariate, tri-variate, and quadrivariate effects. The four-variable parameters depict the degree to which raters agree with each other above the expected agreement given the lower-order log-linear parameters. Therefore, the four-variable parameters represent conditional agreement rates. The (conditional) overall agreement will be high if the four-variable parameters indicating complete agreement are principally high and do not differ from each other, the four-variable parameters indicating disagreement should be low.

If there are special combinations of congruent ratings for two constructs with very high four-variable parameters these categories (of the joint ratings) are associated (lack of discriminant validity). It may be the case that the joint rating of highly extraverted individuals is associated to the joint rating of highly intelligent individuals. In this case, one category of one construct (that is, congruently judged) may serve as an indicator of judgeability for the other construct, the constructs lack of discriminant validity for these categories since their co-occurrence is higher than should be for independent (perfectly discriminant) constructs, or the co-occurrence can be theoretically explained and expected. This has to be examined with respect to the constructs under consideration and with respect to the decision making process. If particular categories of one construct enhance the judgeability on other constructs they should do so for several categories of the other construct and they should do so for several constructs. Then, it is meaningful to conceive this category as an indicator of judgeability. If the category is only associated to one category of one or few other constructs it is very questionable if this particular category indicates the visibility of behavioral cues (good targets sensu Funder, 1995) or if the associated categories represent closely related categories (lack of discriminant validity).

A specific kind of method bias can be examined independently of all other effects examining the log-linear effects of partial agreement. If these are large, this indicates that although peers agree on one construct, they disagree on the other in specific ways. A close examination of the answer process may yield insight into the reasons for the divergent ratings.

ii) Three-variable effects

In models with higher order effects, lower order effects may be interpreted as average effects influencing particular cell combinations. The interpretation of these effects is only meaningful if the higher order effects are absent or have the same qualitative impact (increase or decline of the expected probabilities) on the cells affected by the lower order effect. The same qualitative impact implies that all higher order effects lead to a higher co-occurrence of the category combinations of the lower order effects and the lower order effects may be interpreted as average effects. For sake of simplicity, I consider the case of absent higher order effects.

Assume that all higher order effects are absent. The different *three-variable effects* influence the cells of Table 6.1.2 representing the latent three-variable joint distribution for *XA*, *YS*, and *YA*. However, the implications apply to all possible combinations of three latent variables. These implications can be easily derived reordering the latent variables to follow the patterns presented in Table 6.1.2. That is, first presenting the one variable measuring the 1st (distinct) construct and then the cross-classification of the two latent variables measuring the same construct.

Table 6.1.2

Extracted part of the latent joint distribution in the saturated CT MTMR model for four three-categorical latent variables with three-variable interactions as highest order effects

				YA	
			1	2	3
xa =1	*YS*	1	*A*	M	N
		2	O	*B*	*P*
		3	Q	R	*C*
xa=2	*YS*	1	*G*	1	2
		2	3	*H*	4
		3	5	6	*I*

Note. Only parts of the subtable for three constructs *XA*, *YS*, and *YA* are depicted. The implications account for any other three-variable subtable as well.

Log-linear parameters do not impose any directional link. The effects presented here correspond to correlations and higher order correlations; therefore, it is principally possible to interpret all effects as the influences of any variable on the association of the other two variables. In order to examine rater agreement as a special form of convergent validity it is useful to inspect the meaning of the latent three-variable effects as the influence of one latent construct's score on the joint categorization of the other construct. Therefore, these effects can be interpreted in two principal ways. Three-variable effects either represent properties of judgeable individuals (*A* through *R*) or sources (correlates) of disagreement (an additional form of bias; 1 to 6). These influences are especially meaningful in models when one rater can be conceived as providing better ratings than the other but they may also occur in other cases.

Agreement. The three-variable parameters of cells representing agreement on one construct depict if agreement depends on the category of the other construct. E.g.:

$$\tau^{XA.YS.YA}_{xa.ys.ya} = \frac{\sqrt[XS]{\prod_{w=1}^{XS} \pi^{XS.XA.YS.YA}_{w.xa.ys.ya}}}{\eta * \tau^{XA}_{xa} \tau^{YS}_{ys} \tau^{YA}_{ya} \tau^{XA.YS}_{xa.ys} \tau^{XA.YA}_{xa.ya} \tau^{YS.YA}_{ys.ya}} \text{, with } ys = ya \qquad (6.1.9)$$

indicates to which ratio the geometric mean of all cells belonging to a particular combination of *XA* and identical categories on *YS* and *YA* deviates from what can be expected based on all lower order effects. The following constellations are possible:

- The three-variable parameters of cells representing agreement on one construct are high for specific categories of one variable of the other construct. Then the three-variable effects indicate for which specific categories of *XA* agreement on *Y* is obtained to a higher degree than expected based on the lower order effects. The categories of *XA* can be conceived as a kind of judgeability indicator or as marker categories for good targets. This interpretation is especially meaningful if rater *A* can be conceived as a better rater of the individual's true status than rater *S*. If *S*, for example, correctly judges a target person to be extraverted, *A* and *S* agree more often on their ratings of the target's conscientiousness.

- The three-variable parameters of cells representing agreement on one construct are low for specific categories of the other ones. Then the three-variable effects indicate for which specific categories of *XA* agreement on *Y* is obtained to a smaller degree than expected based on the lower order effects. In this case, specific categories of one construct indicate bad judgeability. In the same vain as highly extraverted individuals may be more easily congruently judged, individuals scoring low on extraversion may not be easily judged on some traits. The three-variable effects, therefore, also may indicate the opposite of judgeability.

- The three-variable parameters of cells representing agreement on one construct are 1 for specific categories of the other ones. Then the three-variable effects indicate that the other construct's category does not have any influence on raters' agreement on the other construct.

Disagreement. The three-variable parameters of cells representing disagreement on one construct depict if this special combination of disagreement is associated to the status on the other construct. E.g.:

$$\tau_{xa.ys.ya}^{XA.YS.YA} = \frac{\sqrt[XS]{\prod_{w=1}^{XS} \pi_{w.xa.ys.ya}^{XS.XA.YS.YA}}}{\eta * \tau_{xa}^{XA} \tau_{ys}^{YS} \tau_{ya}^{YA} \tau_{xa.ys}^{XA.YS} \tau_{xa.ya}^{XA.YA} \tau_{ys.ya}^{YS.YA}}, \text{ with } ys \neq ya \qquad (6.1.10)$$

indicates to which ratio the geometric mean of all cells belonging to a particular combination of *XA* and different categories on *YS* and *YA* deviates from what can be expected based on all lower order effects. The following constellations are possible:

- The three-variable parameters of cells representing disagreement on one construct are high for specific categories of one latent variable of the other construct. The expected proportions are higher for a specific case of disagreement if a particular category is chosen on the other construct. If one of the raters were a better rater and provided ratings that came closer to the true status of an individual this would indicate that the other rater misinterprets behavioral cues (associated to *XS*) leading to a different rating on the other construct (*YA*) although (*YS*) is the better rating. Therefore, this constellation represents rater bias. This may be the case if *S* rates the combination of being highly extraverted (*xs*) and highly neurotic (*ys*) and the other rater *A* does simply not assume highly extraverted to be highly neurotic and therefore only chooses moderately neurotic (*ya*). That is, this effect depicts special cases of higher order rater bias. If the two raters are structurally different but no one is outstanding with respect to the other (no gold-standard rater) this parameter simply indicates differences with respect to the joint ratings. An interpretation of bias is awkward in this case. However, this effect may be interpreted in terms of indicators or behavioral cues that may be ambiguously interpreted by different raters, they differ in the ways they link the behavioral cues to the traits, and indicate on which categories raters disagree enabling researchers to implement new and specific research programs investigating these cell combinations or to train raters.

- The three-variable parameters of cells representing disagreement on one construct are small for specific categories of the other. The expected proportions are smaller for a specific case of disagreement if a particular category is chosen on the other construct. This effect indicates if particular categories of one construct are less often associated (confounded) for a given rating on the other construct. If rater *S* judges the target person to be highly extraverted, this may prevent raters *A* and *S* from providing ratings of not at all neurotic and highly neurotic. This constellation thus indicates to which degree special disagreement combinations do not occur for given statuses on another construct.

- The three-variable parameters of cells representing disagreement on one construct are 1 for specific categories of the other one. Then the three-variable effects indicate that the other construct's category does not have any influence on raters' disagreement on the other construct.

At the level of three-variable parameters, there are some combinations of the above mentioned constellations that merit special attention because these can be related to the concepts of convergent and discriminant validity. One category (say *xs*) can be seen as an indicator of judgeability if this category generally produces higher agreement rates on other constructs (at least on the majority of its categories). The three-variable effect of the same category (*xs*) with disagreement cells indicates if the increase in agreement leads to a decline in disagreement for particular cells or for all cells. That is, if the better judgeability prevents raters from choosing specific category combinations of disagreement or if it prevents them from disagreeing in general. The latter would also automatically lead to an increase of convergent validity.

The three-variable effects indicate a higher order method bias if they are large for cells indicating disagreement on one construct. In this case (only for the given constellation on one rater's ratings), the other rater shows a biased judgment. Bias is understood as the difference between two raters in general (see Agresti, 1992). It is not understood as the difference between a rating and the true score or the true level on a given construct. Method bias type I reflects if the latent prevalence rates differ, the distinguishability index shows if raters distinguish between the categories of one trait, and the method bias introduced here is a conditional distinguishability index showing if one rater deviates from the rating of the other one on the same construct given a particular status on the other construct.

In some cases, one rater (*S*) may be a gold standard providing very accurate ratings. In this case, it is meaningful to mainly inspect the method bias (conditional distinguishability) starting with the category of the gold standard (*xs*) influencing the agreement / disagreement of the ratings on the other construct (*Y*). The three-variable effect of one cell for the non-reference rater (the non-gold-standard method) on the joint rating on the other construct of the reference rater should not be interpreted in this way; but this effect should be interpreted as the influence of the effect of the gold-standards category on the ratings of the non-reference rater. This can be interpreted as a kind of halo effect, which depicts the influence of one trait on the judgments of other traits rated by non-reference raters.

iii) Two- and one-variable effects.

If there are no four- and no three-variable effects the two-variable parameters can be directly interpreted. Their interpretation comes very close to the criteria introduced by Campbell and Fiske (1959).

Assume that all higher order effects (three- and four-variable effects) are absent. The different *two-variable effects* influence the cells of Table 6.1.3 representing the latent two-variable joint distributions for the different bivariate combinations of *XS*, *XA*, *YS*, and *YA*. The upper part [(a), containing the grey-shaded agreement cells] indicates the bivariate distribution of *YS* and *YA* (or *XS* and *XA*, respectively, not depicted). The middle part (b) represents the across trait latent bivariate distribution for *XS* and *YS* (or *XA* and *YA*, not depicted). The lower part (c) represents the across traits - across raters latent bivariate distribution of *XA* and *YS* (or *XS* and *YA*, not depicted).

The latent bivariate sub-tables are completely independent from each other since no three- or four-variable effects are assumed to hold. Therefore, these subtables can be inspected as “complete tables” without any conditional assumption about scores on other variables.

Table 6.1.3

Extracted part of the latent joint distribution in the saturated CT MTMR model with two-variable effects as highest order effects for different combinations of two categorical latent variables

(a)			*YA*	
		1	2	3
YS	1	*A*	1	2
	2	4	*B*	*3*
	3	5	6	*C*

(b)			*YS*	
		1	2	3
XS	1	1	2	3
	2	4	5	6
	3	7	8	9

(c)			*YS*	
		1	2	3
XA	1	10	11	12
	2	13	14	15
	3	16	17	18

Note. Only one pair of variables has been depicted for every kind of association.

Consider the latent subtable representing *monotrait-heteromethod* category combinations, the case of no higher order effects allows for testing the structure of agreement on the level of latent bivariate interactions as described in Section 5.1. Therefore, I will only repeat the main implications of the saturated model here. The structure of agreement is reflected in part (a) of Table 6.1.3.

Method bias type I reflects the degree to which the latent marginals differ from each other. It can be determined as:

$$MB1_{(xs.xa)} = \frac{\pi_{xs}^{XS}}{\pi_{xa}^{XA}}, \text{ for } xs = xa. \tag{6.1.11}$$

It can be analogously determined for the other trait *Y*. Values close to 1 indicate no bias, values far from one indicate over- or underrepresentation of the corresponding latent marginals with respect to the other rater's score (see Agresti, 1992). Values differing from 1 indicate that raters differ with respect to their expected marginals which in turn can be interpreted as different presumed prevalence rates (see Zwick, 1988). This indicates that raters judge the constructs differently. In this sense, method bias is also related to a lack of convergent validity in the log-linear models with latent variables (biased ratings cannot lead to perfect agreement).

Agreement can be seen in high two-variable effects $\left(\tau_{xs.xa}^{XS.XA} \text{ or } \tau_{ys.ya}^{YS.YA}\right)$ for categories of the two trait variables sharing the same index $(xs = xa \text{ or } ys = ya)$. In the special case of a hierarchical model with two-variable parameters as effects of highest order the log-linear two-variable parameters correspond to the category-specific agreement rates. Cells representing agreement (*A*, *B*, and *C*) are grey shaded in Table 6.1.3. An overall latent agreement rate can be calculated using κ.

If there is general (category-specific) agreement beyond agreement on chance at least some disagreement cells are underrepresented. This can be seen in two-variable effects that are smaller than 1 for disagreement cells (1 to 6 in Table 6.1.3 a). The distinguishability index shows which cells are less (more) frequently expected than based on the product of their latent marginals:

$$Dist_{(xs.xa)} = \frac{\pi_{xs.xa}^{XS.XA}}{\pi_{xs}^{XS}\pi_{xa}^{XA}}, \text{ for } xs \neq xa \tag{6.1.12}$$

This index can be analogously defined for the categories of *Y*. If this index is the same for all disagreement cells, raters distinguish equally well between the different categories of the latent constructs and agree more often than predicted by chance.

However, this index can also show values larger than 1 indicating that this particular category combination is more often expected than based on the latent marginals. This indicates that the two raters confound these categories. Or more statistically spoken, the ratings are biased with respect to the other rating. Reconsider the example with the security oriented, gambling, and risk seeking personality types. The two-variable effect indicates to which ratio the 1st rater chooses the gambling personality type if the 2nd rater chooses the risk-seeking personality type. Note that this bias has not to be the same the other way round. That is, the ratio of the combination gambling and risk-seeking personality type does not have to be the same as risk-seeking and gambling personality type.

The association between two latent variables belonging to the same rater but different constructs [part (b) of Table 6.1.3) corresponds to a *heterotrait-monomethod* association sensu Campbell and Fiske (1959):

$$\tau_{xs.ys}^{XS.YS}\text{, or } \tau_{xa.ya}^{XA.YA}. \tag{6.1.13}$$

In general, this effect should be rather weak to indicate discriminant validity. That is, the log-linear two-variable parameters should be close to 1 to indicate discriminant validity. The association between two variables can be category specific. That is, special categories of neuroticism (highly neurotic) may co-occur with particular categories of conscientiousness (moderately conscientious) but not with others. This effect may be due to several (interacting) influences: a theoretical overlap of the categories (a theoretically meaningful category combination; yet, the constructs are not perfectly discriminant), and / or method bias. Method bias is a rater specific view of how categories belonging to two different constructs are related. These effects do not have to be identical across the different raters.

The associations between variables belonging to different constructs judged by different raters [part (c) of Table 6.1.3] correspond to *heterotrait-heteromethod* associations sensu Campbell and Fiske (1959):

$$\tau_{xs.ya}^{XS.YA}\text{, or } \tau_{xa.ys}^{XA.YS}. \tag{6.1.14}$$

These parameters mirror associations between the latent constructs that are shared between raters. These effects can be due to a theoretical overlap of the constructs but they cannot be due to method bias. Therefore, the ratio of the association between traits belonging to one rater (confounded with bias) and the mean association of the corresponding bias free associations indicates the rater specific bias (the rater's view that is, not shared across raters):

Definition 6.1.2

Method bias type II

$$MB2_{(XA.YA)} = \frac{\tau_{xa.ya}^{XA.YA}}{\sqrt{\tau_{xs.ya}^{XS.YA}\tau_{xa.ys}^{XA.YS}}}, \text{ with } xs = xa \text{ and } ys = ya\,. \qquad (6.1.15)$$

This ratio has not yet been defined as method bias to my mind. The denominator gives the expectancy for the bias free association of the latent categories of the two constructs taking the geometric mean of the bias-free associations. The association between the same categories within one method (confounded with bias) is compared to this “average association”. Values larger than 1 indicate an association of the two categories for one rater that goes beyond the bias-free association. That is, one rater implicitly or explicitly associates the two categories to a greater (smaller) extent than do different raters. It reflects rater specific theories or beliefs about the combined prevalences of different statuses (e.g. halo-effect). Values smaller than 1 indicate that this association is less frequently expected than based on the bias-free association - which may be interpreted as an inversed halo-effect. This coefficient is theoretically founded in the postulate of Campbell and Fiske (1959) that the pattern of associations should be the same for all traits in monomethod as well as in heteromethod blocks.

The method bias type II depends on three parameters: The heterotrait-monomethod two-variable interaction and the two heterotrait-heteromethod two-variable interactions representing the same latent categories. Since the denominator is the geometric mean of the two heterotrait-heteromethod parameters this index should not be calculated if the heterotrait-monomethod parameter falls into the interval between the two heterotrait-heteromethod parameters. In this case, the rater-specific view is in the “middle” of the rater-unspecific views; it can therefore not be higher or lower as the error free interaction (if this is conceived as the “average” interaction) and is therefore not biased. Taking the geometric mean of the two heterotrait-heteromethod parameters will most probably lead to a value differing from the numerator implying over- or underrepresentation as the form of bias which is not true inspecting the two-variable effects.

If one of the raters is a gold standard, the method bias type II reduces to the ratio of the heterotrait-monomethod parameter for the non-reference rater to the heterotrait-monomethod parameter of the reference rater (gold standard):

Definition 6.1.3

Method bias type II with gold standard

$$MB2_{(XA.YA)} = \frac{\tau_{xa.ya}^{XA.YA}}{\tau_{xs.ys}^{XS.YS}}\text{, with } xs = xa \text{ and } ys = ya\,. \qquad (6.1.16)$$

if *S* represents a gold standard.

The interpretation of all parameters but the highest-order parameters as presented here can only be done if all higher order effects are absent. However, dealing with empirical data researchers are interested in the agreement rates of their raters. The latent log-linear parameters of lower order effects correspond to “average” effects. Therefore, these effects should only be interpreted (as a directional effect not interpreting the parameter value) if the higher order interactions do not change the direction of the main (lower order) effect for different categories (all parameters of the considered cells must be larger or smaller than 1). The same rationale accounts for the saturated log-linear model.

A heuristic inspection of latent bivariate subtables can be done to get some insight into convergent and discriminant validity sensu Campbell and Fiske (1959). However, if higher order effects are present, the tables are not collapsible. Therefore, I do not recommend inspecting the log-linear parameters of bivariate subtables in cases where higher order effects are present. κ, however may be calculated to get an estimation of general agreement between raters.

6.2 The Correlated Traits Multitrait-Multirater Model for Interchangeable Raters

The saturated model for the analysis of 2 traits by 2 interchangeable raters is a special case of the saturated model for structurally different raters described above. The detailed definition is not repeated but the model equation and the necessary constraints for interchangeable raters are presented. In principle the same logic as in Section 5.2 (latent rater agreement models for interchangeable raters) accounts for the larger 3 x 3 x 3 x 3 model.

6.2.1 Formal Representation of the Saturated CT MTMR Model for Interchangeable Raters

Like in the latent rater agreement models for interchangeable raters measurement invariance has to be assumed. Extending the model to four measurement models (measuring two traits) leads to the following constraints on model parameters defined in Equation 6.1.1[18]:

$$\begin{aligned} e_{\mathbf{abcd}.xa.xb.ya.yb} &= \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \mathrm{T}_{\mathbf{c}} \mathrm{T}_{\mathbf{d}} \tau_{xa}^{XA} \tau_{xb}^{XB} \tau_{ya}^{YA} \tau_{yb}^{YB} \\ &\times \tau_{xa.xb}^{XA.XB} \tau_{xa.ya}^{XA.YA} \tau_{xa.yb}^{XA.YB} \tau_{xb.ya}^{XB.YA} \tau_{xb.yb}^{XB.YB} \tau_{ya.yb}^{YA.YB} \qquad , \\ &\times \tau_{xa.xb.ya}^{XA.XB.YA} \tau_{xa.xb.yb}^{XA.XB.YB} \tau_{xa.ya.yb}^{XA.YA.YB} \tau_{xb.ya.yb}^{XB.YA.YB} \tau_{xa.xb.ya.yb}^{XA.XB.YA.YB} \end{aligned} \tag{6.2.1}$$

with:

$$\max(xa) = \max(xb) = C \text{ and } \max(ya) = \max(yb) = D, \tag{6.2.2}$$

leading to an equal number of categories for the different latent variables representing the same trait, respectively. And:

[18] The variable represented in the model change names from *S* (representing self–report in empirical applications) and *A* to *A* and *B* (representing peer reports *A* and *B*) to prevent from confounding the models. This change does by no means affect the definition or the meaning of the parameters.

$$\tau^{XA}_{xa} = \tau^{XB}_{xb}\text{, for } xa = xb \text{ and } \tau^{YA}_{ya} = \tau^{YB}_{yb}\text{, for } ya = yb, \qquad (6.2.3)$$

the latent distributions are identical for latent variables of different raters measuring the same trait. And:

$$\mathrm{T}_{\mathbf{a}} = \prod_{m_i=1}^{I} \tau^{M_i}_{m_i} \tau^{M_i.XA}_{m_i.xa} = \mathrm{T}_{\mathbf{b}} = \prod_{n_j=1}^{J} \tau^{N_j}_{n_j} \tau^{N_j.XB}_{n_j.xb} \text{ with } \tau^{M_i}_{m_i} = \tau^{N_i}_{n_i} \wedge \tau^{M_i.XA}_{m_i.xa} = \tau^{N_i.XB}_{n_i.xb}, \qquad (6.2.4)$$

and

$$\mathrm{T}_{\mathbf{c}} = \prod_{o_k=1}^{K} \tau^{O_k}_{o_k} \tau^{O_k.YA}_{o_i.ya} = \mathrm{T}_{\mathbf{d}} = \prod_{p_l=1}^{L} \tau^{P_l}_{p_l} \tau^{P_l.YB}_{p_l.yb} \text{ with } \tau^{O_i}_{o_i} = \tau^{P_i}_{p_i} \wedge \tau^{O_i.YA}_{o_i.ya} = \tau^{P_i.YB}_{p_i.yb}, \qquad (6.2.5)$$

indicating identical measurement models. An explanation of these restrictions is given in Section 5.2.

In addition to the restrictions of measurement invariance the interchangeability of the raters has to be respected. The latent more-variable log-linear parameters of the saturated model for interchangeable raters have to be constrained (yielding a symmetry model):

i. $\tau^{XA.XB}_{xa.xb} = \tau^{XA.XB}_{xb.xa}$, (6.2.6)

The log-linear two-variable effects of cells within trait units are identical for inversed ordering of the categories.

ii. $\tau^{XA.YA}_{xa.ya} = \tau^{XB.YB}_{xb.yb}$, (6.2.7)

for $xa = xb \wedge ya = yb$. The rater-specific two-variable effects across constructs (*heterotrait-monomethod* parameters) are the same across raters. The two raters have the same view about which latent categories are related.

iii. $\tau^{XA.YB}_{xa.yb} = \tau^{XB.YA}_{xb.ya}$, (6.2.8) for $xa = xb \wedge ya = yb$. The across traits two-variable effects are the same for the inversed order of raters (the *heterotrait-heteromethod* parameters are identical irrespective of the raters). That is,, if the neurotic category of peer *A* co-occurs more frequently with the conscientious category of *B* this must also be the case (to the same degree) for the opposite combination (neurotic for *B* and conscientious for *A*).

$$\tau_{xa.xb.ya}^{XA.XB.YA} = \tau_{xb.xa.yb}^{XA.XB.YB}, \text{ for } ya = yb, \tag{6.2.9}$$

and

$$\tau_{xa.ya.yb}^{XA.YA.YB} = \tau_{xb.yb.ya}^{XB.YA.YB}, \text{ for } xa = xb. \tag{6.2.10}$$

The impact of one categorical trait variable on the interaction of two categorical trait variables representing the same construct is the same for the two raters. That is, if the combination of *XA* = 2 and *XB* = 3 is more often observed for *YA* = 1 this must also be the case for *XA* = 3 and *XB* = 2 given *YB* = 1.

iv. $$\tau_{xa.xb.ya.yb}^{XA.XB.YA.YB} = \tau_{xb.xa.yb.ya}^{XA.XB.YA.YB}, \tag{6.2.11}$$

That is, interchangeability implies that any given overrepresentation of one specific combination of latent ratings must be the same for the inversed order of the raters. The combination of [2 1 3 1] depends on the same log-linear effect as the combination [1 2 1 3].

6.2.1.1 The impact of the different log-linear effects on the analysis of convergent and discriminant validity

The same considerations about the meaning of lower order effects if higher order interactions are present for the case of structurally different raters account for the case of interchangeable raters. Therefore, the inspection of convergent and discriminant validity is based on the analysis of higher order effects. Convergent and discriminant validity can be analyzed inspecting the complete cross classification of all latent variables.

In Table 6.2.1 36 cells of the latent joint quadrivariate cross-classification (with 81 cells) are depicted. These cells represent agreement, partial agreement, and disagreement cells. The cell entries symbolize the expected cell proportions. I denoted all expected probabilities indicating agreement or partial agreement with capital Latin letters. Identical Latin letters represent identical expected response probabilities. Expected proportions of disagreement cells are denoted using Arabic numbers. Identical numbers identify identical expected cell proportions. As can be easily seen the entries in Table 6.2.1 follow a symmetric scheme. This symmetry is produced by the interchangeability of raters producing identical log-linear parameters. Cells representing agreement with respect to neuroticism *and* conscientiousness are determined by “unique” combinations of log-linear effects (grey shaded and surrounded cells). The expected proportion *A*, for example:

$$A: \quad \begin{aligned} e_{++++1.1.1.1} = \eta * \tau_1^{XA}\tau_1^{XB}\tau_1^{YA}\tau_1^{YB} \\ \times\tau_{1.1}^{XA.XB}\tau_{1.1}^{XA.YA}\tau_{1.1}^{XA.YB}\tau_{1.1}^{XB.YA}\tau_{1.1}^{XB.YB}\tau_{1.1}^{YA.YB} \\ \times\tau_{1.1.1}^{XA.XB.YA}\tau_{1.1.1}^{XA.XB.YB}\tau_{1.1.1}^{XA.YA.YB}\tau_{1.1.1}^{XB.YA.YB} \\ \times\tau_{1.1.1.1}^{XA.XB.YA.YB} \end{aligned}, \qquad (6.2.12)$$

depends on products of effects that do not reappear once in the complete table. The symbols “+” replace the manifest categories. The same is true for expected proportions *B*, *C*, *D*, *E*, and *F*.

Table 6.2.1

Extracted part of the latent joint distribution in the saturated (symmetry) CT MTMR model for interchangeable raters

					YB	
				1	2	3
XA=1	*XB*=1	*YA*	1	*A*	*J*	*K*
			2	*J*	*B*	*L*
			3	*K*	*L*	*C*
	XB=2	*YA*	1	*G*	*1*	*2*
			2	*4*	*H*	*3*
			3	*5*	*6*	*I*
…	…	…	…	…	…	…
XA=2	*XB*=1	*YA*	1	*G*	*4*	*5*
			2	*1*	*H*	*6*
			3	*2*	*3*	*I*
	XB=2	*YA*	1	*D*	*M*	*N*
			2	*M*	*E*	*O*
			3	*N*	*O*	*F*
…	…	…	…	…	…	…

Note. Only the cell combinations for *XA* = 1, X*A* = 2, X*B* =1 and X*B* = 2 are depicted. The scheme applies to all other combinations of latent categories as well (see restrictions i to iv in Section 6.2.1).

The expected proportions of cells representing agreement for only one construct (grey shaded for *Y*) reappear once in the frequency table for inversed categories of the construct upon which the two raters do not agree. These are the frequencies *G*, *H*, and *I* for partial agreement on *Y*, as well as *J*, *K*, and *L* and *M*, *N*, and *O* for partial agreement on *X*. E.g.:

$$\begin{aligned}
& e_{++++1.2.1.1} = \eta * \tau_{1}^{XA}\tau_{2}^{XB}\tau_{1}^{YA}\tau_{1}^{YB} \\
& \qquad \times\tau_{1.2}^{XA.XB}\tau_{1.1}^{XA.YA}\tau_{1.1}^{XA.YB}\tau_{2.1}^{XB.YA}\tau_{2.1}^{XB.YB}\tau_{1.1}^{YA.YB} \\
& \qquad \times\tau_{1.2.1}^{XA.XB.YA}\tau_{1.2.1}^{XA.XB.YB}\tau_{1.1.1}^{XA.YA.YB}\tau_{2.1.1}^{XB.YA.YB} \\
& \qquad \times\tau_{1.2.1.1}^{XA.XB.YA.YB} \\
\text{G:} \qquad & \qquad = \eta * \tau_{2}^{XA}\tau_{1}^{XB}\tau_{1}^{YA}\tau_{1}^{YB} \qquad , \qquad\qquad (6.2.13) \\
& \qquad \times\tau_{2.1}^{XA.XB}\tau_{2.1}^{XA.YA}\tau_{2.1}^{XA.YB}\tau_{1.1}^{XB.YA}\tau_{1.1}^{XB.YB}\tau_{1.1}^{YA.YB} \\
& \qquad \times\tau_{2.1.1}^{XA.XB.YA}\tau_{2.1.1}^{XA.XB.YB}\tau_{2.1.1}^{XA.YA.YB}\tau_{1.1.1}^{XB.YA.YB} \\
& \qquad \times\tau_{2.1.1.1}^{XA.XB.YA.YB} \\
& \qquad = e_{++++2.1.1.1}
\end{aligned}$$

since Equations 6.2.6 through 6.2.11 must hold. All other frequencies also appear two times in the complete table, because they are identical with respect to a complete category inversion. That is, if the latent categories for the two peers are simultaneously interchanged, the model yields the same expected frequency.

The saturated CT MTMR model allows for determining different sources of influences on the associations between latent variables. These coefficients have been defined for the CT MTMR model for structurally different raters. Their meanings with respect to the model for interchangeable raters will be sketched and differences with respect to the model for structurally different raters will be pointed out:

i) Four-variable effects

Complete agreement. The four-variable log-linear parameters of cells indicating agreement on *both* constructs (*A* through *F*) mainly indicate the *judgeability* of the targets. If these effects are larger than 1 and significant, the corresponding expected cell proportions are higher than expected based on all lower order effects:

$$\tau_{xa.xb.ya.yb}^{XA.XB.YA.YB} = \frac{\pi_{xa.xb.ya.yb}^{XA.XB.YA.YB}}{\pi_{xa.xb.ya}^{XA.XB.YA}\pi_{xa.xb.yb}^{XA.XB.YB}\pi_{xa.ya.yb}^{XA.YA.YB}\pi_{xb.ya.yb}^{XB.YA.YB}}, \qquad (6.2.14)$$

for $xa = xb$ and $ya = yb$. Several constellations are possible:

- All four-variable parameters of complete disagreement cells (*A* through *F*) are larger than 1 and of equal size $\tau_{xa.xb.ya.yb}^{XA.XB.YA.YB} = \tau_{(xa.xb.ya.yb)'}^{XA.XB.YA.YB}$ for all $(xa.xb.ya.yb) \neq (xa.xb.ya.yb)'$ with $xa = xb$ and $ya = yb$. This indicates that the interchangeable raters agree constantly across all categories of the different traits. The odds to agree given the expected proportions based on lower order effects (see Eq. 6.1.6) are identical on all category combinations indicating agreement on both constructs.

- All four-variable parameters of complete agreement cells (*A* through *F*) are larger than 1 but differ from each other. In this case, the raters agree more often than expected based on the lower order effects. There is a group of judgeable individuals but their judgeability depends partly on their status on the latent variables. Individuals who belong to an especially easily judgeable category of one trait can be more easily accurately (congruently) judged on a category of the other traits as well. This effect may be weaker or stronger depending on the different categories. $\tau_{xa.xb.ya.yb}^{XA.XB.YA.YB} \leq \tau_{(xa.xb.ya.yb)'}^{XA.XB.YA.YB}$ for all $(xa.xb.ya.yb) \neq (xa.xb.ya.yb)'$ with $xa = xb$ and $ya = yb$. In this case, judgeability (as palpability) is a property of different constellations of the latent categories.

  If there are only few but very large four-variable parameters of complete agreement cells low discriminant validity on agreement ratings is found. The latent categories of the different constructs partly overlap and cannot be considered very distinct from each other.

- All four-variable parameters of complete agreement cells (*A* through *F*) are larger than 1 and differ from each other as a function of categories of one trait. $\tau_{xa.xb.ya.yb}^{XA.XB.YA.YB} \leq \tau_{xa.xb.(ya.yb)'}^{XA.XB.YA.YB}$ for all $(xa.xb.ya.yb) \neq xa.xb.(ya.yb)'$ with $xa = xb$ and $ya = yb$ or $\tau_{xa.xb.ya.yb}^{XA.XB.YA.YB} \leq \tau_{(xa.xb)'.ya.yb}^{XA.XB.YA.YB}$ for all $(xa.xb.ya.yb) \neq (xa.xb)'.ya.yb$ with $xa = xb$ and $ya = yb$. This effect is a special case of the previously described phenomenon. It may occur that for different levels of the target person's extraversion the raters have fewer problems to correctly (at least congruently) classify these targets on the other construct (conscientiousness) with the same accuracy for all categories.

- The four-variable parameters of complete agreement cells (*A* through *F*) may also be smaller than 1. For the corresponding cells agreement on both constructs is less frequently expected than predicted on the lower order effects. This result would be rather awkward but could be explained in cases when the latent cells indicate categories that are (partly) mutually exclusive in the raters' view. Reconsider the example of ratings with respect to clarity of one's own feelings and expressivity of feelings, the cell indicating agreement on “does not show feelings” and “is clear about feelings” can logically be underrepresented because in this case the clarity about feelings is not open for observation.

- The four-variable parameters of complete agreement cells (*A* through *F*) do not differ from 1. In this case the quadrivariate agreement can be explained by lower order effects of agreement (see discussion of these effects below).

Partial agreement. Four-variable parameters of cells indicating agreement on one construct but not on the other for the quadrivariate joint distribution (cells *G* through *O*) represent a special kind of *rater bias*:

$$\begin{aligned}\tau^{XA.XB.YA.YB}_{xa.xb.ya.yb} &= \frac{\pi^{XA.XB.YA.YB}_{xa.xb.ya.yb}}{\pi^{XA.XB.YA}_{xa.xb.ya.}\pi^{XA.XB.YB}_{xa.xb.yb}\pi^{XA.YA.YB}_{xa.ya.yb}\pi^{XB.YA.YB}_{xb.ya.yb}}, \\ &= \tau^{XA.XB.YA.YB}_{xb.xa.yb.ya}\end{aligned} \qquad (6.2.15)$$

for either $(xa = xb \text{ and } ya \neq yb)$ or $(xa \neq xb \text{ and } ya = yb)$. Again, different constellations may occur:

- The four-variable parameters of cells indicating agreement on one but not on the other trait (cells *M* through *O*) are larger than 1. This finding can be interpreted in terms of rater bias. Although raters agree on one construct they disagree systematically on the other construct. Moreover, the particular combination of disagreement cells is equally frequently expected for interchangeable raters.

This effect indicates the association of agreement on one construct with disagreement on the other. Individuals who do not show their feelings (the interchangeable raters agree on the status of “expressivity of feelings”) may, for example, send out ambiguous signals belonging to the construct clarity of feelings which might indicate a very clear or a neutral category. The two interchangeable raters therefore confound these categories with respect to the other rater. This kind of effect may account for all cells indicating partial agreement or only for particular cells.

- The four-variable parameters for (particular or all) cells indicating agreement on one but not on the other trait (cells *M* through *O*) are smaller than 1. In this case, disagreement between the two raters with respect to specific category combinations is underrepresented if they agree on the other construct. This may be the case if agreement on one construct is very hard to achieve because the trait under consideration is not easily judgeable, if two raters agree on judging this difficult trait, they will most probably agree on more easy to judge traits as well and therefore the expected proportions of the disagreement cells for the latter construct are much smaller given agreement on the first trait.

  This effect thus shows (if there is agreement) that there is higher agreement on one construct (on all or on one category) if there is agreement on the other one. The opposite does not necessarily have to be true. In this case, one construct (or specific cells of this construct) is more difficult to judge than the cells of the other construct.

- All four-variable parameters for cells indicating agreement on one but not on the other trait (cells *M* through *O*) do not differ from 1. In this case, agreement on one construct is not related to disagreement on the other construct.

Disagreement. The latent four-variable parameters of cells besides the agreement and partial agreement cells represent influences which may be due to bias or to general disagreement:

$$\begin{aligned}\tau^{XA.XB.YA.YB}_{xa.xb.ya.yb} &= \frac{\pi^{XA.XB.YA.YB}_{xa.xb.ya.yb}}{\pi^{XA.XB.YA}_{xa.xb.ya}\pi^{XA.XB.YB}_{xa.xb.yb}\pi^{XA.YA.YB}_{xa.ya.yb}\pi^{XB.YA.YB}_{xb.ya.yb}}, \\ &= \tau^{XA.XB.YA.YB}_{xb.xa.yb.ya}\end{aligned} \qquad (6.2.16)$$

for $(xa \neq xb \text{ and } ya \neq yb)$. The following different constellations are possible:

- All four-variable parameters of complete disagreement cells (1 to 6) are larger than 1. In this case, the two raters disagree more often than predicted based on the lower order effects. In general, this indicates a lack of convergent validity. κ will be negative. However, there still might be a few positive category-specific agreement ratios for some cells. I do not expect this constellation to appear in any application. This constellation may appear in cases where raters do not follow their instructions or due to a wrong labeling of categories. Even if raters are guessing they should have four-variable parameters for disagreement cells that do not differ from 1.

- Some (one) four-variable parameters of complete disagreement cells (1 to 6) are larger than 1. In this case particular combinations of one rater's latent scores are associated to the other rater's scores but for different cell combinations. If raters weigh some behavioral cues differently given cues on the other trait they may be more often categorized in latent disagreement cells. If, for example, one rater classifies an individual due to specific behavioral cues as highly extraverted and, additionally, these cues may lead this rater to also classify this individual as moderately neurotic this combination of behavioral cues may be associated to the moderately extraverted and highly neurotic classes for the other rater. The same effect has to hold for inversed categories across raters (the opposite combination).

- Some (all) four-variable parameters of complete disagreement cells (1 to 6) are smaller than 1. This may be due to higher complete and / or partial agreement rates. Higher agreement lowers the expected proportions of the disagreement cells in the saturated model. Yet, this may also be due to high disagreement on a particular cell combination and no effects on complete or partial agreement cells.

- None of the four-variable parameters of complete disagreement cells (1 to 6) differs from 1. In this case, there is neither an over- nor an underrepresentation of complete disagreement cells

At the level of four-variable effects, there are some combinations that can be related to the concepts of convergent and discriminant validity. In principle, these relations do not differ from those for structurally different raters except for the interchangeability of the raters.

Overall agreement may be high due to bivariate, tri-variate, and quadrivariate effects. The four-variable parameters depict the degree to which raters agree with each other above the expected agreement given the lower-order log-linear parameters. Therefore, the four-variable parameters represent conditional agreement rates. The (conditional) overall agreement will be high if the four-variable parameters indicating complete agreement are principally high and do not differ from each other, the four-variable parameters indicating disagreement should be low.

If there are special combinations of congruent ratings for two constructs with very high four-variable parameters these categories (of the joint ratings) co-occur more often than expected based on the lower order effects (lack of discriminant validity). It may be the case that the joint rating of highly extraverted individuals co-occurs with the joint rating of highly intelligent individuals. In this case, one category of one construct may serve as an indicator of judgeability for the other construct, the constructs lack of discriminant validity for these categories, or this effect can be theoretically explained and expected. This has to be examined with respect to the constructs under consideration. If particular categories of one construct enhance the judgeability on other constructs they should do so for several categories of the other construct and they should do so for several constructs. Then, it is meaningful to conceive this category as an indicator of judgeability. If the category is only associated to one category of one or few other constructs it is very questionable if this particular category indicates if individuals are judgeable (good targets sensu Funder, 1995) or if the categories represent closely related categories (lack of discriminant validity).

A specific kind of method bias can be examined independently of all other effects examining the log-linear effects of partial agreement. If these are large, this indicates that although peers agree on one construct, they confound categories of the other construct in specific (and inversely related) ways. A close examination of the answer process and the category definition may yield insight into the reasons for this kind of method bias (which corresponds to a kind of “category confusion”).

Method bias is strong when the latent four-variable effects influencing cells indicating disagreement are large and the four-variable effects influencing (complete and partial) agreement are small. The joint ratings are associated but not with respect to agreement.

ii) Three-variable effects

In general, lower order effects may be interpreted as average effects influencing particular cell combinations. The interpretation of these effects is only straightforward if the higher order effects are absent or all higher order effects influencing the cells of that particular lower order effect increase (or decline) the expected cell proportions.

Assume that all higher order effects are absent. Table 6.2.2 represents parts of the latent three-variable joint distributions for combinations of *XA* or *XB* with *YA* and *YB*. However, the implications account for every possible combination of three variables and can be easily derived reordering the latent variables to follow the patterns presented in Table 6.2.2. The symmetric structure is the same as in Table 6.2.1. Latin letters again indicate agreement on one construct (on *Y* in Table 6.2.2) and Arabic numbers indicate disagreement (on *Y* in Table 6.2.2).

In order to examine rater agreement as a special form of convergent validity it is useful to inspect the meaning of the latent three-variable effects as the influence of one latent construct's score on the joint categorization of the other construct. Therefore, these effects can be interpreted in two principal ways. Three-variable effects either represent properties of judgeable individuals (*A* through *F*) or sources of disagreement (1 to 12).

These effects have to be identical across the interchangeable raters. This does imply that the rate of expected classification on one trait for a given constellation on the other trait increases or declines to the same degree for interchangeable raters, however, it does not say that the raters congruently choose the same category (this is depicted in the four-variable effects) but that congruent ratings on the 1st trait variable are only related to chance if there is agreement on the 2nd trait variable.

Table 6.2.2

Extracted part of the latent joint distribution in the saturated CT MTMR model with three-variable effects as highest order effects for four three-categorical latent variables

				YB	
			1	2	3
XB=1	*YA*	1	*A*	1	2
		2	4	*B*	3
		3	5	6	*C*
XB=2	*YA*	1	*D*	7	8
		2	10	*E*	9
		3	11	12	*F*
…	…	…	…	…	…
XA=1	*YA*	1	*A*	4	5
		2	1	*B*	6
		3	2	3	*C*
XA=2	*YA*	1	*D*	10	11
		2	7	*E*	12
		3	8	9	*F*
…	…	…	…	…	…

Note. Only parts of the subtables for the constructs *XA*, *XB*, *YA*, and *YB* are depicted. The implications account for any other three-variable subtable as well.

Agreement. The three-variable parameters of cells representing agreement on one construct depict if agreement depends on the status of the other construct. E.g.:

$$\tau_{xa.ya.yb}^{XA.YA.YB} = \frac{\sqrt[XB]{\prod_{xb=1}^{XB} \pi_{xa.xb.ys.yb}^{XA.XB.YA.YB}}}{\eta * \tau_{xa}^{XA} \tau_{ya}^{YA} \tau_{yb}^{YB} \tau_{xa.ya}^{XA.YA} \tau_{xa.yb}^{XA.YB} \tau_{ya.yb}^{YA.YB}} \text{, with } ya = yb \qquad (6.2.17)$$

indicates to which ratio the geometric mean of all cells belonging to a particular combination of *XA* and identical categories on *YA* and *YB* deviates from what can be expected based on all lower order effects. The following constellations are possible:

- The three-variable parameters of cells representing agreement on one construct are high for specific categories of the other one. Then the three-variable effects indicate for which specific categories of *XA* agreement on *Y* is obtained to a higher degree than expected based on the lower order effects. The categories of *XA* can be conceived as a kind of judgeability indicator. If one of the raters identifies the target individual to belong to a category indicating judgeability, the raters will more often agree with each other.

- The three-variable parameters of cells representing agreement on one construct are low for specific categories of the other one. Then the three-variable effects indicate for which specific categories of *XA* agreement on *Y* is obtained to a smaller degree than expected based on the lower order effects. In this case, specific categories of one construct indicate bad judgeability.

- The three-variable parameters of cells representing agreement on one construct are 1 for specific categories of the other one. Then the three-variable effects indicate that the other construct's category does not have any influence on raters' agreement on the other construct.

Disagreement. The three-variable parameters of cells representing disagreement on one construct depict if this cell combination is associated to the status on the other construct (cells 1 to 12 in Table 6.2.2). E.g.:

$$\tau^{XA.YA.YB}_{xa.ya.yb} = \frac{\sqrt[XB]{\prod_{xb=1}^{XB} \pi^{XA.XB.YA.YB}_{xa.xb.ya.yb}}}{\eta * \tau^{XA}_{xa} \tau^{YA}_{ya} \tau^{YB}_{yb} \tau^{XA.YA}_{xa.ya} \tau^{XA.YB}_{xa.yb} \tau^{YA.YB}_{ya.yb}}, \text{ with } ya \neq yb \qquad (6.2.18)$$

indicates to which ratio the geometric mean of all cells belonging a particular combination of *XA* and different categories on *YA* and *YB* deviates from what can be expected based on all lower order effects. The following constellations are possible:

- The three-variable parameters of cells representing disagreement on one construct are high for specific categories of the other. This effect is more easily interpreted as the association of a specific combination of one rater's joint classification with the classification of the other rater on one construct. This effect indicates that, for example, if *A* judges the target person to be highly neurotic and moderately conscientious *B* will judge the same target to be moderately neurotic. The same association must hold for the inversed combination (*B* judges highly neurotic and moderately conscientious while *A* judges moderately neurotic). This effect thus reveals easily confounded category constellations.

- The three-variable parameters of cells representing disagreement on one construct are small for specific categories of the other. The expected proportions are smaller for a specific case of disagreement if a particular category is chosen on the other construct. This effect indicates if particular categories of one construct co-occur less often than predicted based on the lower order effects for a given rating on the other construct. If rater *A* judges the target person to be highly extraverted, this may prevent raters *A* and *B* from providing ratings of not at all neurotic and highly neurotic. This constellation thus indicates to which degree special disagreement combinations do not occur for given statuses on another construct. That is, if some latent categories of one trait moderate the disagreement on the other trait (i.e., prevent from misinterpreting behavioral cues or make behavioral cues of the other trait more salient).

- The three-variable parameters of cells representing disagreement on one construct are 1 for specific categories of the other one. Then the three-variable effects indicate that the other construct's category does not have any influence on raters' disagreement on the other construct.

iii) Two- and one-variable effects.

If there are no four- and no three-variable effects the two-variable effects can be directly interpreted. Their interpretation comes very close to the criteria introduced by Campbell and Fiske (1959).

For sake of simplicity, assume that all higher order effects (three- and four-variable effects) are absent. The different *two-variable effects* influence the cells of Table 6.2.3 representing the latent two-variable joint distributions for the different combinations of *XA*, *XB*, *YA*, and *YB*. The upper part [(a), containing the grey-shaded agreement cells] indicates the bivariate distribution of *YA* and *YB* (or *XA* and *XB*, respectively, not depicted). The middle part (b) represents the across trait latent bivariate distribution for *XA* and *YA* (or *XB* and *YB*, not depicted). The lower part (c) represents the across traits-across raters latent bivariate distribution of *XA* and *YB* (or *XB* and *YA*, not depicted).

The latent bivariate sub-tables are completely independent from each other since no three- or four-variable effects are assumed to hold. Therefore, these subtables can be inspected as “complete tables” without any conditional assumption about scores on other variables.

Table 6.2.3

Extracted part of the latent joint distribution in the saturated CT MTMR model with two-variable effects as highest order effects for different combinations of two categorical latent variables

(a)			*YB*	
		1	2	3
YA	1	*A*	1	2
	2	1	*B*	*3*
	3	2	3	*C*

(b)			*YA*	
		1	2	3
XA	1	4	5	6
	2	5	7	8
	3	6	8	9

(c)			*YB*	
		1	2	3
XA	1	10	11	12
	2	11	13	14
	3	12	14	15

Note. Only one pair of variables has been depicted for every kind of association.

I will first consider the latent subtable representing *monotrait-heteromethod* category combinations. The case of no higher order effects allows for testing the structure of agreement on the level of latent bivariate interactions as described in Section 5.2. Therefore I will only repeat the main implications of the saturated model here. The structure of agreement is reflected in part (a) of Table 6.2.3.

Method bias type I would reflect the degree to which the latent marginals differ from each other. Since these are restricted to be identical across interchangeable raters this bias does not occur in models with interchangeable raters

Agreement can be seen in high two-variable effects $\left(\tau_{xa.xb}^{XA.XB} \text{ or } \tau_{ya.yb}^{YA.YB}\right)$ for categories of the two trait variables sharing the same index $\left(xa = xb \text{ or } ya = yb\right)$. In the special case of a model with two-variable effects as effects of highest order the log-linear two-variable parameters correspond to the category-specific agreement rates. Cells representing agreement (*A*, *B*, and *C*) are grey shaded in Table 6.2.3. An overall latent agreement rate can be calculated using κ.

If there is general (category-specific) agreement beyond agreement on chance at least some disagreement cells are underrepresented. This can be seen in two-variable effects that are smaller than 1 for disagreement cells (1 to 3 in Table 6.2.3). The two-variable effects show which cells are less frequently expected than based on the product of their latent marginals:

$$\frac{\pi_{xa.xb}^{XA.XB}}{\pi_{xa}^{XA}\pi_{xb}^{XB}} = \tau_{xa.xb}^{XA.XB}, \text{ for } xa \neq xb \qquad (6.2.19)$$

This index can be analogously defined for the categories of *Y*. If this index is smaller than one and the same for all disagreement cells, raters distinguish equally well between the different categories of the latent constructs and agree more often than predicted by chance.

However, this index can also show values larger than 1 indicating that particular category combinations are more often expected than based on the latent marginals. This indicates that the two raters confound these categories. The ratings are biased with respect to the other rating. Reconsider the example with the security oriented, gambling, and risk seeking personality types. The two-variable effect indicates to which ratio the 1st rater chooses the gambling personality type if the 2nd rater chooses the risk-seeking personality type. Note that this bias has to be the same the other way round for interchangeable raters. That is, the ratio of the combination gambling and risk-seeking personality type is the same as risk-seeking and gambling personality type.

The association between two latent variables belonging to the same rater but different constructs [part (b) of Table 6.2.3] corresponds to a *heterotrait-monomethod* association sensu Campbell and Fiske (1959):

$$\tau^{XA.YA}_{xa.ya} = \frac{\pi^{XA.YA}_{xa.ya}}{\pi^{XA}_{xa}\pi^{YA}_{ya}}, \text{ or } \tau^{XB.YB}_{xb.yb} = \frac{\pi^{XB.YB}_{xb.yb}}{\pi^{XB}_{xb}\pi^{YB}_{yb}}. \tag{6.2.20}$$

In general, these effects should be rather weak to indicate discriminant validity. That is, the log-linear two-variable parameters should be close to 1 to indicate discriminant validity. This effect may be due to several (interacting) influences: a theoretical overlap of the categories (theoretically meaningful overrepresentation of the joint category; yet, the constructs are not perfectly discriminant), and / or method bias. Method bias is a rater specific view of associations between categories belonging to two different constructs. These associations must be identical across the different raters.

The associations between variables belonging to different constructs judged by different raters [part (c) of Table 6.1.3] correspond to *heterotrait-heteromethod* associations sensu Campbell and Fiske (1959):

$$\tau^{XA.YB}_{xa.yb} = \frac{\pi^{XA.YB}_{xa.yb}}{\pi^{XA}_{xa}\pi^{YB}_{yb}}, \text{ or } \tau^{XB.YA}_{xb.ya} = \frac{\pi^{XB.YA}_{xb.ya}}{\pi^{XB}_{xb}\pi^{YA}_{ya}}. \tag{6.2.21}$$

These parameters mirror interactions between the latent categories across raters. These effects can be due to a theoretical overlap of the constructs but they cannot be due to method bias. Therefore, method bias type II can be estimated in the models for interchangeable raters: the ratio of the association between traits belonging to one rater (confounded with bias) and the mean association of the corresponding bias free associations indicates the rater specific bias (the rater's view that is, not shared across raters):

$$MB2_{(XA.YA)} = \frac{\tau^{XA.YA}_{xa.ya}}{\sqrt{\tau^{XA.YB}_{xa.ya}\tau^{XB.YA}_{xa.ya}}} = \frac{\tau^{XA.YA}_{xa.ya}}{\tau^{XB.YA}_{xa.ya}} = \frac{\tau^{XA.YA}_{xa.ya}}{\tau^{XA.YB}_{xa.ya}} = \frac{\tau^{XB.YB}_{xa.ya}}{\sqrt{\tau^{XA.YB}_{xa.ya}\tau^{XB.YA}_{xa.ya}}} = MB2_{(XB.YB)}. \tag{6.2.22}$$

This ratio of the joint classification across traits belonging to one rater (confounded with bias) and the mean joint classification of the corresponding bias free associations indicates the rater specific bias (the rater's view that is, not shared across raters). This bias is the same for the two interchangeable raters. This bias can always be determined because the heterotrait-heteromethod associations do not differ for interchangeable raters.

The interpretation of all parameters but the highest-order parameters in their pure forms as presented here can only be done if all higher order effects are absent. However, dealing with empirical data researchers are interested in the agreement rates of their raters. The latent log-linear parameters of lower order effects correspond to “average” effects. Therefore, these effects should only be interpreted (as a directional effect not interpreting the parameter value) if the higher order interactions do not change the direction of the main (lower order) effect for different categories. A heuristic inspection of latent bivariate subtables can be done to get some insight into convergent and discriminant validity sensu Campbell and Fiske (1959). However, if higher order effects are present, the tables are not collapsible. Therefore, I do not recommend inspecting the log-linear parameters of bivariate subtables in cases where higher order effects are present. κ, however may be calculated to get an estimation of general agreement between raters.

6.3 Empirical Applications of the CT MTMR Model for Structurally Different and Interchangeable Raters

In this section, the CT MTMR models for structurally different and for interchangeable raters will be applied to the empirical data described in Section 4.1.3. First, the models for structurally different raters analyzing the combination of self-report and peer report *A* data will be reported and illustrated, then the model for interchangeable raters analyzing the two peer reports *A* and *B* will be applied

The computationally very complex CT MTMR models are prone to several problems during the estimation process: sparse table problems leading to meaningless *p*-values of the χ^2-parameters, boundary solutions due to intrinsic or empirical model non-identification, and zero fitted marginals or cell frequencies (which also lead to boundary values and undefined log-linear parameters). Therefore, researchers should absolutely check the results obtained from one program against different start-values and cross-validate their results using different statistical packages. However, to date, there is no other program than LEM allowing (at least in parts) for these complex analyses. Therefore, all model parameters and interpretations of these should only be considered as illustrative. The model results will be discussed with respect to their expected proportions because almost all log-linear parameters suffer from boundary values.

6.3.1 Empirical Applications of the CT MTMR Model for Structurally Different Raters

The CT MTMR model for structurally different raters will be applied to the self-report and peer report *A* data measuring neuroticism and conscientiousness. The most complex model allowing for all two-, three-, and four-variable effects will be presented first. In two steps the four- and three-variable effects will be removed.

Table 6.3.1

Goodness-of-fit coefficients of the CT MTMR models for structurally different raters

Highest Effects	χ^2	$p\left(\chi^2\right)$	L^2	$p(L^2)$	*df*	AIC[1]	BIC[1]
4	79125224.54	.00	6418.70	1.00	43046544	-86086669	-265574001
3	69473766.35	.00	6425.23	1.00	43046560	-86086694	65574093
2	79696342.45	.00	6479.59	1.00	43046592	-86086704	-265574236

Note. Highest Effects: 4, 3, and 2 indicate the four-, three-, and two-variable effects as highest order effects in the models. χ^2: Pearson χ^2-value; L^2 likelihood-based χ^2-value; [1]AIC and BIC are based on L^2-values; the bootstrap is not available for these models due to memory size restrictions in the DOS routine of LEM.

Table 6.3.1 presents the goodness-of-fit criteria for the different models. However, during the estimation process the following problems occurred: LEM is known to have difficulties estimating the standard errors for models with more than 150 parameters[19]. Therefore, no information on boundary values can be determined for the model with four-variable effects (176 parameters) and for the model with three-variable effects (160

[19] http://spitswww.uvt.nl/web/fsw/mto/lem/lembugs.txt

parameters) as highest-order effects. Moreover, the bootstrap routine in LEM did not work for any model (for structurally different or interchangeable raters) due to memory restrictions of the DOS routine.

The model with two-variable interactions as highest-order interactions consists of 128 parameters. For this model 7 parameters near the boundaries of the parameter space were found. Inspecting the outputs for the models with higher-order interactions (four- and three- variable interactions) reveals that all log-linear parameters representing effects of latent variables are extremely large or very close to 0. However, the measurement models can be soundly estimated and do not differ with respect to the applications in Sections 4 and 5. Therefore, I will exemplify the impact of the higher-order interactions relying on the latent expected probabilities for these models.

6.3.1.1 Results of the CT MTMR model with four-variable interactions for structurally different raters

Table 6.3.2 depicts the quadrivariate latent joint distribution of the cross classification of the latent variable representing neuroticism and conscientiousness rated by a self-rater and peer rater *A*. The model equation for the population is:

$$\begin{aligned} e_{\mathbf{abcd}.ns.na.cs.ca} = {} & \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \mathrm{T}_{\mathbf{c}} \mathrm{T}_{\mathbf{d}} \\ & \times \tau_{ns}^{NEUS} \tau_{na}^{NEUA} \tau_{cs}^{CONS} \tau_{ca}^{CONA} \\ & \times \tau_{ns.na}^{NEUS.NEUA} \tau_{ns.cs}^{NEUS.CONS} \tau_{ns.ca}^{NEUS.CONA} \tau_{na.cs}^{NEUA.CONS} \tau_{na.ca}^{NEUA.CONA} \tau_{cs.ca}^{CONS.CONA} , \\ & \times \tau_{ns.na.cs}^{NEUS.NEUA.CONS} \tau_{ns.na.ca}^{NEUS.NEUA.CONA} \tau_{ns.cs.ca}^{NEUS.CONS.CONA} \tau_{na.cs.ca}^{NEUA.CONS.CONA} \\ & \times \tau_{ns.na.cs.ca}^{NEUS.NEUA.CONS.CONA} \end{aligned} \qquad (6.3.1)$$

with *ns* and *na* representing the latent categories of the latent trait variables *NEUS* and *NEUA* for self-rated (*S*) and peer rated neuroticism and *cs* and *ca* representing the latent categories of the latent trait variables *CONS* and *CONA* for self-rated (*S*) and peer rated (*A*) conscientiousness.

Table 6.3.2

Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with four-variable effects for structurally different raters

			CONA					
			1		2		3	
NEUS=1	*NEUA*=1	*CONS*=1	.01	(.00)	.01	(.00)	.01	(.00)
		CONS=2	**.02**	**(.00)**	**.03**	**(.00)**	.00	(.00)
		CONS=3	.00	(.00)	.01	(.00)	**.02**	**(.00)**
	NEUA=2	*CONS*=1	.00	(.00)	.00	(.00)	.01	(.00)
		CONS=2	.00	(.00)	.00	(.00)	.01	(.00)
		CONS=3	.00	(.00)	.00	(.00)	**.04**	**(.00)**
	NEUA=3	*CONS*=1	.00	(.00)	.01	(.00)	.00	(.00)
		CONS=2	.00	(.00)	.01	(.00)	.01	(.00)
		CONS=3	.00	(.00)	.00	(.00)	**.03**	**(.00)**
NEUS=2	*NEUA*=1	*CONS*=1	**.02**	**(.00)**	**.02**	**(.00)**	.00	(.00)
		CONS=2	.00	(.00)	.00	(.00)	.00	(.00)
		CONS=3	.00	(.00)	.00	(.00)	.01	(.00)
	NEUA=2	*CONS*=1	.01	(.00)	.00	(.00)	**.02**	**(.00)**
		CONS=2	.01	(.00)	**.03**	**(.00)**	**.03**	**(.01)**
		CONS=3	.00	(.00)	**.02**	**(.00)**	**.07**	**(.01)**
	NEUA=3	*CONS*=1	.01	(.00)	.00	(.00)	.00	(.00)
		CONS=2	.01	(.00)	**.02**	**(.00)**	**.05**	**(.01)**
		CONS=3	.00	(.00)	**.02**	**(.00)**	**.05**	**(.01)**
NEUS=3	*NEUA*=1	*CONS*=1	.01	(.00)	.01	(.00)	.00	(.00)
		CONS=2	.01	(.00)	.00	(.00)	.01	(.00)
		CONS=3	.00	(.00)	.01	(.00)	.01	(.00)
	NEUA=2	*CONS*=1	.00	(.00)	.01	(.00)	.01	(.00)
		CONS=2	.00	(.00)	**.02**	**(.00)**	.01	(.00)
		CONS=3	.00	(.00)	.00	(.00)	**.04**	**(.00)**
	NEUA=3	*CONS*=1	.01	(.00)	**.04**	**(.00)**	.01	(.01)
		CONS=2	**.02**	**(.00)**	**.02**	**(.00)**	.02	(.01)
		CONS=3	.00	(.00)	.01	(.00)	**.06**	**(.01)**

Note. Entries in bold type depict expected proportions that deviate from the predictions based on the marginals by more than one decimal. Entries in parentheses represent the product of the latent marginals.

The log-linear parameters cannot be soundly estimated in LEM and thus these parameters cannot be interpreted in terms of over- or underrepresentations for particular latent cells of the joint distribution. Calculating the odds and odds ratios in order to inspect if latent quadrivariate or trivariate cells are over- or underrepresented does not solve the problem because divisions by 0 occur (this is a necessary consequence of boundary solutions). However, the latent joint distribution can be heuristically examined revealing expected proportions that are more frequently expected than given the product of their latent marginals. Expected proportions that differ to an extent of 2% or more from the product of their marginals (depicted in parentheses) are printed in bold type. This inspection of the expected cell proportions can only be heuristic compared to an inspection of (not available) properly estimated log-linear parameters and their standard errors. The log-linear parameters could be used to identify effects and their impact on the latent joint distribution and to test them statistically.

In total 22 bold typed entries can be found in Table 6.3.2. That is, in 22 out of 81 cells comparably high expected proportions can be found. In a first step, cells indicating overall agreement $\left(\hat{\pi}_{ns.na.cs.ca}^{NEUS.NEUA.CONS.CONA}, \text{ with } ns = na \text{ and } cs = ca\right)$ will be considered. These 9 cells are principally expected more often than could be predicted by the product of their latent marginals. About 26% of the latent ratings can be found on the overall agreement diagonal. 6 out of the 9 cells are printed in bold type. That is, there is considerable agreement on both constructs at the same time. This agreement more often occurs for cell combinations with high conscientious $(cs = ca = 3)$ individuals being either sensitive but stable $(ns = na = 2)$ or highly neurotic $(ns = na = 3)$. 18% out of the 26% of agreement can be found in these cells (that is, 69% of the overall agreement cells fall into these combinations).

Raters may also agree with respect to one construct but disagree with respect to the other one (partial agreement). Raters agree 27% of the time on their ratings for conscientiousness when they disagree with respect to neuroticism. This leads to an overall agreement on conscientiousness of 53% (for the complete table). 7 cells indicating partial agreement on conscientiousness differ to an extent of 2% or more from the product of their latent marginals. The main proportion of the agreements on conscientiousness can be found for moderately or highly conscientious individuals. Peers seem to have difficulties judging a not conscientious individual congruently with the self-rater on this trait. With respect to the self-raters as reference raters the peer rating is biased for not conscientious individuals.

For the latent construct of neuroticism raters agree 50% of the time in total. About half of the time, they agree with respect to neuroticism they also agree with respect to conscientiousness (26%, see above). Agreement on neuroticism is higher for individuals being sensitive but stable or highly neurotic. 79% (19% of 24%) of the partial agreement fall into these cells. 6 cells indicating partial agreement on neuroticism differ to an extent of 2% or more from the product of their latent marginals. Agreement on neuroticism can be found to a greater extent for higher scores on this variable (being neurotic or sensitive but stable). This finding is in line with other findings that traited individuals can be more congruently rated (see Baumeister & Tice, 1988; Funder, 1995 for an overview).

Disagreement cells do not differ to a large extent from what is predicted by the product of their latent marginals. The only combinations that are more frequently expected are cells for the combinations of being sensitive but stable in the self-report $\left(ns = 2\right)$ and highly neurotic in the peer report $\left(na = 3\right)$ with being moderately conscientious rated by the self- or peer rater $\left(cs = 3 \vee ca = 3\right)$ and / or moderately conscientious by the self- or peer rater $\left(cs = 2 \vee ca = 2\right)$.

Peer raters who agree with the targets that the targets are highly neurotic do not agree with them if targets indicate not to be conscientious but judge them to be moderately conscientious. The same is true for agreement on being sensitive but stable. In this case, self-ratings indicating not to be conscientious are associated to a high peer-perceived level of conscientiousness$\left(\hat{\pi}_{2.2.1.3}^{NEUS.NEUA.CONS.CONA} = .02\right)$. Conscientiousness and neuroticism seem to be related for moderate or high scores on neuroticism at least in the peer view.

It is important to note, that these analyses are carried out by inspecting the table of expected frequencies consisting of 81 cells. Much better information could be gained by an inspection of log-linear parameters which identify the underlying effects of the different expected proportions. The high overall agreement rate implies that there is an association between agreement on one construct and agreement on the other construct, but without a statistical test it remains unclear if the corresponding overrepresentation is due to a four-variable effect, emerges from lower order effects, or even is a random association. The same is true for the associations concerning the disagreement cells. A comparison of the quadrivariate latent joint distribution to the one implied by the model with 2nd order interactions as interactions of highest order may (heuristically) give more insight into the question if 3rd order interaction are present.

6.3.1.2 Results of the CT MTMR model with three-variable interactions as highest order interactions for structurally different raters

Table 6.3.3 depicts the quadrivariate latent joint distribution of the latent trait variables for neuroticism and conscientiousness rated by a self-rater and peer rater *A*. The model equation for the population is:

$$\begin{aligned} e_{\mathbf{abcd}.na.nb.ca.cb} &= \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \mathrm{T}_{\mathbf{c}} \mathrm{T}_{\mathbf{d}} \\ &\times \tau_{ns}^{NEUS} \tau_{na}^{NEUA} \tau_{cs}^{CONS} \tau_{ca}^{CONA} \\ &\times \tau_{ns.na}^{NEUS.NEUA} \tau_{ns.cs}^{NEUS.CONS} \tau_{ns.ca}^{NEUS.CONA} \tau_{na.cs}^{NEUA.CONS} \tau_{na.ca}^{NEUA.CONA} \tau_{cs.ca}^{CONS.CONA}, \\ &\times \tau_{ns.na.cs}^{NEUS.NEUA.CONS} \tau_{ns.na.ca}^{NEUS.NEUA.CONA} \tau_{ns.cs.ca}^{NEUS.CONS.CONA} \tau_{na.cs.ca}^{NEUA.CONS.CONA} \end{aligned} \qquad (6.3.2)$$

with *ns* and *na* representing the latent categories of the latent trait variables *NEUS* and *NEUA* for self-rated (S) and peer rated neuroticism and *cs* and *ca* representing the latent categories of the latent trait variables *CONS* and *CONA* for self-rated (*S*) and peer rated (*A*) conscientiousness. Expected proportions that differ to an extent of 2% or more from the product of their marginals (depicted in parentheses) are printed in bold type.

In total 20 bold typed entries can be found in Table 6.3.3. That is, in 20 out of 81 cells comparably high expected proportions can be found. The 9 cells indicating overall agreement $\left(\hat{\pi}_{ns.na.cs.ca}^{NEUS.NEUA.CONS.CONA}, \text{with } ns = na \text{ and } cs = ca\right)$ are principally expected more often than could be predicted by the product of their latent marginals. About 28% of the latent ratings can be found on the overall agreement diagonal. 6 out of the 9 cells are printed in bold type. That is, there is considerable agreement on both constructs at the same time. This agreement more often occurs for cell combinations with high conscientious $(cs = ca = 3)$ individuals being either sensitive but stable $(ns = na = 2)$ or highly neurotic $(ns = na = 3)$. 25% out of the 28% of agreement can be found in these cells (that is, 89% of the overall agreement cells fall into these combinations).

Table 6.3.3

Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with three-variable effects as highest order interactions for structurally different raters

			CONA					
			1		2		3	
NEUS=1	*NEUA*=1	*CONS*=1	.01	(.00)	**.02**	**(.00)**	.00	(.00)
		CONS=2	**.03**	**(.00)**	**.03**	**(.00)**	.00	(.00)
		CONS=3	.00	(.00)	.01	(.00)	**.02**	**(.00)**
	NEUA=2	*CONS*=1	.00	(.00)	.00	(.00)	.01	(.00)
		CONS=2	.00	(.00)	.00	(.00)	.01	(.00)
		CONS=3	.00	(.00)	.00	(.00)	**.04**	**(.00)**
	NEUA=3	*CONS*=1	.00	(.00)	.01	(.00)	.00	(.00)
		CONS=2	.00	(.00)	.01	(.00)	.01	(.00)
		CONS=3	.00	(.00)	.00	(.00)	.03	(.00)
NEUS=2	*NEUA*=1	*CONS*=1	**.02**	**(.00)**	**.02**	**(.00)**	.00	(.00)
		CONS=2	.00	(.00)	.00	(.00)	.00	(.00)
		CONS=3	.00	(.00)	.00	(.00)	.01	(.00)
	NEUA=2	*CONS*=1	.01	(.00)	.00	(.00)	**.02**	**(.00)**
		CONS=2	.01	(.00)	**.04**	**(.00)**	.02	(.01)
		CONS=3	.01	(.00)	.02	(.00)	**.07**	**(.01)**
	NEUA=3	*CONS*=1	.01	(.00)	.00	(.00)	.00	(.01)
		CONS=2	.01	(.00)	.01	(.00)	.05	(.01)
		CONS=3	.00	(.00)	.01	(.00)	.05	(.01)
NEUS=3	*NEUA*=1	*CONS*=1	.01	(.00)	**.02**	**(.00)**	.00	(.00)
		CONS=2	.01	(.00)	.00	(.00)	.01	(.00)
		CONS=3	.01	(.00)	.01	(.00)	.01	(.00)
	NEUA=2	*CONS*=1	.00	(.00)	.00	(.00)	.01	(.00)
		CONS=2	.00	(.00)	**.02**	**(.00)**	.01	(.00)
		CONS=3	.00	(.00)	.00	(.00)	**.04**	**(.00)**
	NEUA=3	*CONS*=1	.01	(.00)	**.04**	**(.01)**	.01	(.01)
		CONS=2	**.02**	**(.00)**	**.03**	**(.01)**	.02	(.01)
		CONS=3	.00	(.00)	.01	(.01)	**.06**	**(.01)**

Note. Entries in bold type depict expected proportions that deviate from the predictions based on the marginals by more than one decimal. Entries in parentheses represent the product of the latent marginals.

Raters may also agree with respect to one construct but disagree with respect to the other one (partial agreement). Raters agree 26% of the time on their ratings for conscientiousness when they disagree with respect to neuroticism. This leads to an overall agreement on conscientiousness of 54% (for the complete table). 6 cells indicating partial agreement on conscientiousness differ to an extent of 2% or more from the product of their latent marginals. The main proportion of agreement on conscientiousness can be found for moderately or highly conscientious individuals. Peers seem to have difficulties judging a not conscientious individual congruently with the self-rater on this trait. With respect to the self-raters as reference raters the peer rating is biased for not conscientious individuals.

For the latent construct of neuroticism raters agree 52% of the time in total. About half of the time, they agree with respect to neuroticism they also agree with respect to conscientiousness (28% see above). Agreement on neuroticism is higher for individuals being sensitive but stable or highly neurotic. 75% (18% of 24%) of the partial agreement fall into these cells. 5 cells indicating partial agreement on neuroticism differ to an extent of 2% or more from the product of their latent marginals.

Disagreement cells do not differ to a large extent from what is predicted by the product of their latent marginals. The only combination that is more frequently expected are cells for the combinations of being not conscientious in the self-report (*cs* = 1) and moderately conscientious in peer report *A* (*ca* = 2) for targets that have been judged not neurotic by *A* for all statuses of self-reported neuroticism. Additionally, there principally is agreement between self- and peer raters concerning low conscientiousness. Therefore, one may conclude that peers deviate from the self-reported score on conscientiousness for low self-rated conscientious individuals if peers perceive the target person as not neurotic.

Peer raters who agree with the targets that the targets are highly neurotic do not agree with them if targets indicate not to be conscientious but judge them to be moderately conscientious $\left(\hat{\pi}_{3.3.1.2}^{NEUS.NEUA.CONS.CONA}=.04\right)$. The same is true for agreement on being sensitive but stable, in this case, self-ratings indicating not to be conscientious are associated to a high peer perceived level of conscientiousness $\left(\hat{\pi}_{2.2.1.3}^{NEUS.NEUA.CONS.CONA}=.02\right)$. Conscientiousness and neuroticism seem to be related for moderate or high scores on neuroticism at least in the peer view.

It is important to emphasize that these interpretations have been carried out relying on expected proportions and not on the comparison of log-linear effects with corresponding standard errors. Therefore, all interpretations can only be considered illustrative.

6.3.1.3 Results of the CT MTMR model with two-variable effects as highest order interactions for structurally different raters

Table 6.3.4 depicts the quadrivariate latent joint distribution of the cross classification of the latent variables representing neuroticism and conscientiousness rated by a self-rater and peer rater *A*. The model equation for the population is:

$$\begin{aligned} e_{\mathbf{abcd}.na.nb.ca.cb} &= \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \mathrm{T}_{\mathbf{c}} \mathrm{T}_{\mathbf{d}} \\ &\times \tau_{ns}^{NEUS} \tau_{na}^{NEUA} \tau_{cs}^{CONS} \tau_{ca}^{CONA} \\ &\times \tau_{ns.na}^{NEUS.NEUA} \tau_{ns.cs}^{NEUS.CONS} \tau_{ns.ca}^{NEUS.CONA} \tau_{na.cs}^{NEUA.CONS} \tau_{na.ca}^{NEUA.CONA} \tau_{cs.ca}^{CONS.CONA} \end{aligned} , \qquad (6.3.3)$$

with *ns* and *na* representing the latent categories of the latent trait variables *NEUS* and *NEUA* for self-rated (*S*) and peer rated (*A*) neuroticism and *cs* and *ca* representing the latent categories of the latent trait variables *CONS* and *CONA* for self-rated (*S*) and peer rated (*A*) conscientiousness.

In contrast to the two previously described models, the log-linear parameters of the model with only two-variable interactions can be interpreted. However, in order to make the interpretation of the model comparable to the other models (and to the model for interchangeable raters, see below) the expected proportions are presented, the log-linear parameters are presented in Appendix E. The interpretation of these parameters corresponds to the conclusion drawn from the expected proportions (and is, therefore, redundant). Since the boundary values afflicting the log-linear parameters cannot be considered a priori model parameters the *z*-values provided by LEM cannot be interpreted (see Galindo-Garre & Vermunt, 2004, 2005, 2006).

Table 6.3.4

Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with two-variable effects as highest order interactions for structurally different raters

			CONA					
			1		2		3	
NEUS=1	*NEUA*=1	*CONS*=1	**.02**	**(.00)**	**.02**	**(.00)**	.00	(.00)
		CONS=2	.01	(.00)	**.02**	**(.00)**	.01	(.00)
		CONS=3	.00	(.00)	.01	(.00)	**.02**	**(.00)**
	NEUA=2	*CONS*=1	.00	(.00)	.01	(.00)	.01	(.00)
		CONS=2	.00	(.00)	.01	(.00)	**.02**	**(.00)**
		CONS=3	.00	(.00)	.00	(.00)	**.04**	**(.00)**
	NEUA=3	*CONS*=1	.00	(.00)	.00	(.00)	.00	(.00)
		CONS=2	.00	(.00)	.01	(.00)	.01	(.00)
		CONS=3	.00	(.00)	.00	(.00)	.01	(.00)
NEUS=2	*NEUA*=1	*CONS*=1	.01	(.00)	.01	(.00)	.00	(.00)
		CONS=2	.01	(.00)	.01	(.00)	.00	(.00)
		CONS=3	.00	(.00)	.01	(.00)	.01	(.00)
	NEUA=2	*CONS*=1	.01	(.00)	.01	(.00)	.01	(.00)
		CONS=2	.01	(.00)	**.02**	**(.00)**	**.04**	**(.01)**
		CONS=3	.00	(.00)	.01	(.00)	**.09**	**(.01)**
	NEUA=3	*CONS*=1	.01	(.00)	.01	(.00)	.01	(.00)
		CONS=2	.01	(.00)	**.02**	**(.00)**	**.02**	**(.00)**
		CONS=3	.00	(.00)	.01	(.00)	**.05**	**(.01)**
NEUS=3	*NEUA*=1	*CONS*=1	.01	(.00)	**.02**	**(.00)**	.00	(.00)
		CONS=2	.01	(.00)	.01	(.00)	.00	(.00)
		CONS=3	.00	(.00)	.01	(.00)	.01	(.00)
	NEUA=2	*CONS*=1	.00	(.00)	.01	(.00)	.01	(.00)
		CONS=2	.00	(.00)	.01	(.00)	**.02**	**(.00)**
		CONS=3	.00	(.00)	.01	(.00)	**.04**	**(.00)**
	NEUA=3	*CONS*=1	.01	(.00)	**.02**	**(.00)**	.01	(.00)
		CONS=2	.01	(.00)	**.03**	**(.00)**	**.02**	**(.00)**
		CONS=3	.00	(.00)	.01	(.00)	**.05**	**(.00)**

Note. Entries in bold type depict expected proportions that deviate from the predictions based on the marginals by more than one decimal. Entries in parentheses represent the product of the latent marginals.

Table 6.3.4 depicts the expected proportions of the quadrivariate latent joint distribution. As before, expected cell proportions that differ for at least 2% from the product of the latent marginals are depicted in bold type. In total the entries of 19 cells are bold typed. 7 out of these 19 represent overall agreement. 8 cells represent partial agreement and 4 represent total disagreement. The reduction in the number of expected proportions that deviate from the product of their marginals can be explained by the more restrictive form of this model. The interplay between the latent variables is much more restricted than in the models presented before.

Overall agreement cells comprise about 27% of the sample, the highest entries can be found for the agreement combinations of highly conscientious with either sensitive but stable or neurotic personality types (14% of all entries fall into these two joint categories). The agreement rates are principally higher for individuals who are at least moderately conscientious and at least sensitive but stable.

Partial agreement for conscientiousness (26%) can mostly be found for highly conscientious individuals (16% of the joint judgments). For neuroticism a similar pattern can be found 15% out of the 20% of the partial agreement can be found for sensitive but stable or neurotic individuals. Overall, the heuristic analyses inspecting the expected proportions do not differ between the three models (with different levels of interactions).

A more thorough insight into the interplay of the four latent variables can be gained inspecting the bivariate latent distributions. In non-saturated hierarchical models, the joint distributions of the variables corresponding to the highest order interactions are exactly reproduced.

Table 6.3.5 presents the latent rater agreement sub-model for neuroticism. In order to compare the two model implied latent marginal distributions with each other, the method bias type I can be determined:

$$\begin{aligned} &MB1_{(ns=1.na=1)} = \frac{.25}{.25} = 1.00 \\ &MB1_{(ns=2.na=2)} = 1.03 \qquad . \\ &MB1_{(ns=3.na=3)} = 0.98 \end{aligned} \tag{6.3.4}$$

This index shows that the two raters yield ratings with almost perfectly the same prevalence rates. This is a prerequisite for high agreement (see Zwick, 1988).

Table 6.3.5

Cross-classification of expected proportions for the latent variables representing neuroticism

	NEUA			
	1	2	3	
ns = 1	.12 (.06)	.09 (.10)	.04 (.09)	.25
ns = 2	.06 (.11)	.22 (.17)	.14 (.14)	.42
ns = 3	.07 (.08)	.10 (.14)	.16 (11)	.33
	.25	.41	.34	

Note. Values in parentheses represent the product of the latent marginals.

Inspecting the cells on the main diagonal shows considerable agreement. The category-specific agreement rates are in the range of 1.27 to 1.92 (see Table 6.3.6) with the highest value for the latent cell combination of not being neurotic. This finding could not be expected with respect to the quadrivariate latent distribution. However, due to the very small expected proportions in this cell, even small absolute agreement rates will produce large effects. These effects are comparable to the monotrait-heteromethod effects sensu Campbell and Fiske (1959). $\kappa = .24$ indicates a relatively low agreement between the raters.

Table 6.3.6

Distinguishability index and category-specific agreement rates for neuroticism

	NEUA		
	1	2	3
ns = 1	1.92	0.88	0.47
ns = 2	0.57	1.27	0.99
ns = 3	0.86	0.75	1.40

The inspection of the disagreement cells besides the main diagonal also shows an interesting pattern. Table 6.3.5 depicts their expected proportions and the expected values given the latent marginals. All but one cell [2 1] show lower expected proportions than would be expected based on the latent marginals. The distinguishability indices in Table 6.3.6 reflect this finding in a standardized way:

$$Dist_{(x.y)} = \frac{\pi_{x.y}^{X.Y}}{\pi_x^X \pi_y^Y}, \text{ for } x \neq y. \qquad (5.1.3, \text{ repeated})$$

It can be seen that the cell combinations [2 1] and [1 3] are only about half as often expected as predicted by the marginals. Self-rated not neurotic individuals are rarely judged to be neurotic by the peer rater $\left(Dist_{(1.3)} = 0.47\right)$. In the same vain, sensitive but stable self-rated individuals are less often rated not neurotic$\left(Dist_{(2.1)} = 0.57\right)$. Peers obviously perceive if individuals are sensitive (self-rated). They also do not overestimate the self-rated neuroticism score producing no overestimation for the combination of sensitive but stable for the self-report and neurotic for the peer report $\left(Dist_{(2.3)} = 0.99\right)$, however, peers also do not distinguish between these categories. All other distinguishability indices show that self- and peer raters show lower disagreement, yet, they do not differ vastly from the product of the latent marginals (absolutely and relatively). Self-raters and peers discriminate fairly well between the different categories of neuroticism.

Table 6.3.7

Cross-classification of expected proportions for the latent variables representing conscientiousness

	CONA			
	1	2	3	
cs = 1	.08 (.04)	.10 (.07)	.06 (.12)	.24
cs = 2	.07 (.06)	.14 (.11)	.15 (.18)	.35
cs = 3	.02 (.07)	.07 (.13)	.32 (.21)	.41
	.17	.31	.52	

Note. Values in parentheses represent the product of the latent marginals.

Table 6.3.7 depicts the latent bivariate distribution of the latent variables representing conscientiousness. Calculating the method bias type I coefficient:

$$\begin{aligned} &MB1_{(cs=1.ca=1)} = 1.41 \\ &MB1_{(cs=2.ca=2)} = 1.13, \\ &MB1_{(cs=3.ca=3)} = 0.79 \end{aligned} \tag{6.3.5}$$

reveals that self- and peer raters deviate considerably in their latent marginals. Peers rate the targets in more than half of the times as highly conscientious (1.27 times more often than the self-raters). Self-raters choose the lower categories more often. This finding may be due to the fact that the targets are almost exclusively students. In order to successfully complete one’s studies a specific level of conscientiousness is required, peers may attribute the fact that targets complete their work as students to their personality whereas the self-raters may compare themselves to others and do not perceive themselves as conscientious. Moreover, they know about their own possible difficulties in completing the work (e.g., procrastination) and therefore rate themselves lower on conscientiousness. In terms of the rater accuracy model (Funder, 1995), one might conclude that better (more diverse) information is needed for the peer raters to achieve higher agreement rates.

The entries on the main diagonal also show agreement of the two raters with respect to conscientiousness (high convergent validity). The category-specific agreement rates (depicted on the main diagonal of Table 6.3.8) indicate that the overrepresentation is in the range of 1.25 to 2.03. Again the overrepresentation for the lowest category is highest. $\kappa = .28$ indicates relatively low overall agreement.

Disagreement is higher for the cell combinations of moderately conscientious and not conscientious in both ways. That is, self- and peer raters confound these categories to some extent. However, for the disagreement cells with highly conscientious ratings there is no confusion at all. Being highly conscientious on either rating prevents from being classified as moderately or not conscientious. Traited individuals (in the sense of having a high score on a trait) can thus be rated without confusion (Baumeister & Tice, 1988). For conscientiousness, self- and peer raters discriminate well for traited individuals and poorly for moderately and low traited individuals.

The pattern of disagreement differs from what has been found for neuroticism. If targets are not traited this leads to some confusion for conscientiousness, if they are traited this leads to less confusion and higher agreement for conscientiousness. Agreement is principally higher for neuroticism but there is no confusion for individuals being not neurotic. This illustrates that moderators of agreement (Funder, 1995) may have differential impacts with respect to the trait under consideration.

Table 6.3.8

Distinguishability index and category-specific agreement rates for conscientiousness

	CONA		
	1	2	3
cs = 1	2.03	1.37	0.44
cs = 2	1.12	1.25	0.81
cs = 3	0.30	0.56	1.49

Table 6.3.9

Cross-classification of expected proportions for the latent variables originating in the self-report

	CONS			
	1	2	3	
ns = 1	.08 (.07)	.09 (.09)	.09 (.11)	.26
ns = 2	.08 (.11)	.16 (.15)	.18 (.17)	.42
ns = 3	.10 (.08)	.10 (.12)	.13 (.14)	.33
	.25	.35	.41	

Note. Values in parentheses represent the product of the latent marginals.

Table 6.3.9 depicts the latent joint classification of the trait variables originating in the self-report. Obviously, there is little deviation from the expected proportions and the product of the latent marginals. This indicates that the self-raters distinguish well between the two latent traits. For self-raters, these traits are not associated (see also Section 4.1.4). This indicates almost perfect discriminant validity sensu Campbell and Fiske (1959).

Table 6.3.10

Cross-classification of expected proportions for the latent variables originating in the peer report

	CONA			
	1	2	3	
na = 1	.08 (.04)	.10 (.08)	.07 (.13)	.25
na = 2	.03 (.07)	.10 (.13)	.28 (.21)	.41
na = 3	.05 (.06)	.12 (.11)	.18 (.18)	.34
	.17	.31	.52	

Note. Values in parentheses represent the product of the latent marginals.

Table 6.3.10 presents the cross-classification of the latent trait variables originating in the peer report. Peers perceive the two constructs as related rating other individuals. To their mind the combination of not being neurotic and highly conscientious appears less often than predicted based on the marginals. Peers rather tend to choose the 1st categories on both variables. Additionally, they perceive sensitive but stable individuals as highly conscientious and less frequently as not conscientious or moderately conscientious. Therefore, one may conclude that there is a lack of discriminant validity with respect to these two traits for peer ratings. However, this lack only concerns particular categories and does not generalize across all possible constellations because the other combinations do not deviate to a great extent from the product of their marginals. It would be interesting to examine if this peer-specific view is linked to a naïve theory on which categories can be related or if this is due to a misinterpretation or detection of behavioral cues leading peers to show associated ratings of neuroticism and conscientiousness. These are question related to the rater accuracy model (Funder, 1995).

Table 6.3.11

Cross-classification of expected proportions for neuroticism originating in the self-report and conscientiousness originating in the peer report

	CONA			
	1	2	3	
ns = 1	.05 (.04)	.07 (.08)	.12 (.13)	.25
ns = 2	.07 (.07)	.12 (.13)	.24 (.21)	.42
ns = 3	.05 (.06)	.12 (.11)	.16 (.17)	.33
	.17	.31	.52	

Note. Values in parentheses represent the product of the latent marginals.

Table 6.3.11 depicts the latent cross-classification of the self-rated neuroticism-scores and the peer rated conscientiousness-scores. There is virtually no deviation from the product of the latent marginals. Self- and peer ratings of the different traits are completely distinct from each other. This indicates high discriminant validity across raters. If one considers the self-rating as a better approximation of the truth peers do not erroneously interpret neurotic behaviors as conscientious.

Table 6.3.12
Cross-classification of expected proportions for neuroticism originating in the peer and conscientiousness originating in the self-report

	CONS			
	1	2	3	
na = 1	.09 (.06)	.09 (.09)	.07 (.10)	.25
na = 2	.08 (.10)	.14 (.14)	.20 (.17)	.41
na = 3	.07 (.09)	.13 (.12)	.19 (.14)	.34
	.24	.35	.41	

Note. Values in parentheses represent the product of the latent marginals.

Table 6.3.12 presents the latent cross-classification of the peer rated neuroticism-scores and the self-rated conscientiousness-scores. The two latent trait variables (*NEUA* and *CONS*) are associated to a stronger degree than the previously presented trait variables. If self-raters perceive themselves as highly conscientious peers do no longer tend to judge them not neurotic but choose the middle and high category of neuroticism. That is, high conscientiousness is slightly confounded with neuroticism in the peer view if one considers the self-rater as better raters than the peers.

Table 6.3.13
Method bias type II for the self-report

	CONS		
	1	2	3
ns = 1	—	1.07	—
ns = 2	1.21	1.26	0.85
ns = 3	1.62	0.84	0.75

Note. — indicates that *MB*2 is meaningless in this cell.

Considering that no rater is outstanding with respect to the other rater allows determining the method bias type II as in Definition 6.1.2:

$$MB2_{ns.cs} = \frac{\hat{\pi}_{ns.cs}^{NEUS.CONS}}{\sqrt{\hat{\pi}_{ns.ca}^{NEUS.CONA}\hat{\pi}_{na.cs}^{NEUA.CONS}}}. \qquad (6.3.6)$$

Tables 6.3.13 and 6.3.14 present the method bias type II parameters. Empty cells indicate indices that are meaningless since the monomethod association is in the range of the two heteromethod associations. The method bias type II for equally good raters is presented in order to have the more general presentation. If one rater is considered to be a better rater than the other one the method bias type II coefficient of Definition 6.1.3 can be calculated. Table 6.3.13 reveals that self-raters tend to rate themselves as highly neurotic but not conscientious, sensitive but stable (middle category) but not conscientious, and sensitive but stable and moderately conscientious more often than on average.

The self-raters conceive themselves less frequently as highly neurotic combined with highly conscientious or moderately conscientious than predicted by the average ratings. The same is true for sensitive but stable individuals who perceive themselves not as often as highly conscientious as predicted by the joint ratings.

Table 6.3.14
Method bias type II for the peer report

	CONA		
	1	2	3
na = 1	–	1.26	0.70
na = 2	0.48	0.75	1.29
na = 3	0.80	0.94	—

Note. — indicates that *MB*2 is meaningless in this cell.

A completely different picture is given by Table 6.3.14 for the bias of peer ratings. The combinations of not neurotic and not conscientious as well as highly neurotic and highly conscientious are not biased with respect to the joint ratings. A positive bias (overrepresentation) can be found for the combinations of not neurotic and moderately conscientious and sensitive but stable (middle category) and highly neurotic. All other cells are less frequently expected than predicted by the joint ratings. Peers do not associate low conscientiousness to the latent statuses being sensitive but stable or neurotic as expected by the average association. The same is true for the combination of not neurotic and highly conscientious.

Self-raters and peers thus differ with respect to the cells that are over- or underrepresented in the cross-classification of their latent variables. Peers perceive the targets principally as more conscientious (see method bias type I) than do self-raters. Additionally, they show larger expected frequencies for two particular cell-combinations of the latent traits. That is, sensitive but stable individuals (middle category) are rated more often as highly conscientious compared to the self-ratings and not neurotic individuals are rated more often as moderately conscientious. These combinations are not overrepresented in the self-report. Therefore, these coefficients reflect a view that is specific to the peer raters. In the same vain, the peers show underrepresentations of the cells for not conscientious ratings and sensitive but stable (middle category) and neurotic individuals. Again, this underrepresentation is specific to the view of peers because self-raters show overrepresented ratings for these categories. The two raters also differ with respect of their views concerning the association of targets being moderately conscientious and sensitive but stable. While self-raters choose this category combination more often than could be expected relying on the bias-free associations (between raters) peers tend to underestimate this association.

6.3.1.4 Summary of the findings for the CT MTMR models for structurally different raters

The applications of the CT MTMR model showed (as expected) that the estimation of complex models with several latent variables is a tedious task and computational very demanding. The models with higher-order interactions yield many boundary values and aberrant parameter estimates. A possible remedy for this problem could emerge from newly developed estimation algorithms shortly mentioned in Section 4.1.2. However, to date these procedures cannot be used in conjunction with the complex models I developed for structurally different raters. Therefore, the models with four- and three-variable interactions as highest order effects can only heuristically be analyzed.

Yet, the model with two-variable interactions as highest order effects yields (relatively) sound parameter estimates which allow for analyzing the latent bivariate relationships. An inspection of the log-linear parameters is redundant (see Appendix E) concerning the associations of the variables and does not provide information about the significance of model parameters because boundary solutions were encountered preventing from inspecting the significance of the effects (Galindo-Garre & Vermunt, 2004, 2005, 2006).

The analysis of the model with two-variable interactions as highest order effects showed some interesting results with respect to the convergent and discriminant validity, method bias, and accuracy of the different raters. There is a considerable overall agreement rate showing that in about 1 out of 4 cases self- and peer raters agree with respect to both constructs. Inspecting the expected proportions may lead to the conclusion that agreement is highest for cell combinations of highly conscientious and sensitive but stable and neurotic individuals. The partial agreement rates also show that self- and peer raters agree more often for individuals classified in one of the above mentioned categories. These findings are in line with the findings for the CT MTMR models with four- and three-variable interactions.

Since the CT MTMR model with two-variable interactions is a hierarchical model it “reproduces” the latent bivariate joint distributions allowing for a direct interpretation of the expected bivariate proportions and the latent one-variable marginals. The method bias type I reveals if the latent marginal distributions differ from each other. This is not the case for neuroticism but for conscientiousness. Peers overestimate the conscientiousness with respect to the self-ratings.

The category-specific agreement rates can be calculated to identify the overrepresentation in the cells on the (agreement) main diagonals. There is agreement for all cells on the two bivariate main diagonals (for neuroticism and conscientiousness). In the model with two-variable interactions as highest order interactions, these effects are the same in all subtables given the categories of the other variables (no four- and three-variable interactions). The category-specific agreement rates show that there is a much higher agreement for the combinations of the lowest categories for both traits. However, this does not imply that these rates are absolutely very high but high with respect to what can be expected knowing the marginals. If there were a possibility to estimate the saturated model with correct standard errors one might inspect the corresponding log-linear effects to judge the ratio of agreement more exactly (note that these effects will differ from subtable to subtable if the higher order effects are present).

The distinguishability index reveals in a similar manner as the category-specific agreement rates if disagreement cells are over- or underrepresented. This index can be used to detect sources of disagreement. For conscientiousness this index revealed, for example, that self- and peer raters confound the first two categories (lack of distinguishability). All other categories can be relatively well distinguished from each other for the two traits (except for the combination of sensitive but stable in the self-report and neurotic in the peer report, which has an expected proportion as predicted by chance). This finding (if replicated and soundly estimated) might serve as a starting point to investigate the decision making process concerning these categories in more depth.

The cross-classification of trait-variables belonging to the same method (heterotrait-monomethod associations) showed that there are virtually no associations for the self-report. However, the peer ratings were associated to some degree revealing that their view about personality types (combinations of latent categories) differs from the self-raters' view. Comparing these associations (for both raters) to the average association of the across raters (heterotrait-heteromethod) association yields the method bias type II. This index shows that self- and peer raters differ with respect to the categories of the two traits they choose. If the self-rating is considered to be a better approximation of the “true-scores” on the two trait variables a comparison of the peer reported classifications to the self-rated classification could be used as method bias type II index.

The two tables representing the heterotrait-heteromethod associations (Table 6.3.11 and 6.3.12) indicate the rater bias free associations between the two traits. These associations are rather weak indicating high discriminant validity. The only cell combination that is constantly slightly overrepresented is the combination of sensitive but stable (middle category) and highly neurotic.

In sum, I conclude that the CT MTMR model could be used to detect category-specific sources of convergence, category-specific lack of discriminant validity as well as distinguishability, allows for a comparison of within raters associations across traits to estimate the rater-specific biases, and (theoretically) to examine if higher agreement rates are due to two-, three-, and / or four-variable effects, that is, if there are moderators of agreement (convergent validity). These pieces of information go far beyond the pieces of information researchers can retrieve of the models presented in Section 2 and of the latent rater agreement models. However, since there is no sound estimation procedure yet, the interpretation of all model parameters must remain heuristic.

6.3.2 Applications of the CT MTMR Models for Interchangeable Raters

The applications of the CT MTMR models for interchangeable raters suffer from some specific problems in the estimation algorithm implemented in LEM. LEM is known to bug for large models with more than 150 parameters and to sometimes produce incorrect results when equality restrictions are implemented in the model definition[20]. These two points account for the saturated models with four-variable interactions as highest order effects. Dropping the four-variable interaction does not remedy the problem although in this case the number of parameters is reduced from 164 to 130. Neither the log-linear parameters nor the expected cell proportions could be estimated according to the model definition. This makes clear that new estimation methods and more advanced programs are needed to soundly estimate the models.

6.3.2.1 Results of the CT MTMR model with two-variable effects as highest order effects for interchangeable raters

The goodness-of-fit indices for the model with two-variable interactions as highest order interactions show divergent results for the different χ^2-values. The Pearson χ^2-value indicates bad fit to the data $\left(\chi^2 = 133083603.94\text{ ; } df = 43046642; p = .00\right)$, the likelihood-ratio based χ^2-value indicates perfect fit to the data $\left(L^2 = 133083603.94\text{ ; } df = 43046642; p = 1.00\right)$. Unfortunately, the bootstrap DOS-routine does not work due to memory restrictions. The AIC and BIC indices $\left(AIC = -86086640; BIC = -265574380\right)$ may serve for model comparison but are meaningless in themselves to assess the goodness-of-fit. The 78 log-linear parameters

[20] http://spitswww.uvt.nl/web/fsw/mto/lem/lembugs.txt

suffer from 12 boundary solutions. Moreover, the model estimation did not work relying on the effect coding scheme but had to be carried out using dummy coding (see Appendix E). Therefore, the log-linear parameters cannot be interpreted as in the model definition. Among others Hagenaars (1990) explains how to interpret dummy-coded log-linear parameters. However, the expected proportions can be interpreted as before.

Tables 6.3.15 and 6.3.16 depict the conditional response probabilities for neuroticism and conscientiousness implied by the CT MTMR model with two-variable effects as highest order interactions for interchangeable raters. The model equation for the population is:

$$\begin{aligned} e_{\mathbf{abcd}.na.nb.ca.cb} = {} & \eta \mathrm{T}_{\mathbf{a}} \mathrm{T}_{\mathbf{b}} \mathrm{T}_{\mathbf{c}} \mathrm{T}_{\mathbf{d}} \\ & \times \tau_{na}^{NEUA} \tau_{nb}^{NEUB} \tau_{ca}^{CONA} \tau_{cb}^{CONB} \\ & \times \tau_{na.nb}^{NEUA.NEUB} \tau_{na.ca}^{NEUA.CONA} \tau_{na.cb}^{NEUA.CONB} \tau_{nb.ca}^{NEUB.CONA} \tau_{nb.cb}^{NEUB.CONB} \tau_{ca.cb}^{CONA.CONB} , \\ & \times \tau_{na.nb.ca}^{NEUA.NEUB.CONA} \tau_{na.nb.cb}^{NEUA.NEUB.CONB} \tau_{na.ca.cb}^{NEUA.CONA.CONB} \tau_{nb.ca.cb}^{NEUB.CONA.CONB} \\ & \times \tau_{na.nb.ca.cb}^{NEUA.NEUB.CONA.CONB} \end{aligned} \qquad (6.3.7)$$

with *na* and *nb* representing the different categories of the latent trait variables *NEUA* and *NEUB* (neuroticism rated by peer *A* or peer *B*) as well as *ca* and *cb* representing the different categories of the latent trait variables *CONA* and *CONB* (conscientiousness rated by peer *A* or peer *B*). $\mathrm{T}_{\mathbf{a}}, \mathrm{T}_{\mathbf{b}}, \mathrm{T}_{\mathbf{c}}$, and $\mathrm{T}_{\mathbf{d}}$ represent the measurement models of the four different TMUs.

Table 6.3.15

Conditional response probabilities for neuroticism in the CT MTMR model with two-variable effects as highest order interactions for interchangeable raters

		latent variable		
Variable	manifest categories	*na* = *nb* = 1	*na* = *nb* = 2	*na* = *nb* = 3
I / Q (vulnerable)	1	.51	.01	.03
	2	.42	.45	.03
	3	.07	.54	.94
J / R (sensitive)	1	.63	.09	.04
	2	.33	.57	.06
	3	.04	.34	.90
K / S (moody)	1	.71	.68	.33
	2	.20	.24	.29
	3	.09	.08	.38
L / T (doubtful)	1	.72	.48	.22
	2	.19	.33	.27
	3	.10	.19	.51

The conditional response probabilities are restricted to be identical across raters within traits according to the implications for interchangeable raters. Empirically, they do not deviate from the conditional response probabilities found for the previously presented models with interchangeable peer raters (see Section 6.2) and for the conditional response probabilities found for peer ratings (*A*) in models with structurally different raters. Therefore, I do not repeat the detailed analyses here. The same implications hold with respect to the latent variables. That is, the three categories of neuroticism are not neurotic targets, sensitive but stable targets (middle category), and neurotic targets. The three categories for conscientiousness range from low to moderately and highly conscientious.

Table 6.3.16

Conditional response probabilities for conscientiousness in the CT MTMR model with two-variable effects as highest order interactions for interchangeable raters

Variable	manifest categories	latent variable *ca* = *cb* =1	 *ca* = *cb* =2	 *ca* = *cb* =3
M / U (industrious)	1	.84	.07	.00
	2	.15	.72	.04
	3	.01	.21	.96
N / V (diligent)	1	.93	.08	.00
	2	.05	.79	.04
	3	.02	.12	.96
O / W (dutiful)	1	.49	.09	.01
	2	.34	.41	.07
	3	.18	.51	.92
P/ X (ambitious)	1	.51	.01	.03
	2	.42	.45	.03
	3	.07	.54	.94

Table 6.3.17 depicts the quadrivariate latent joint distribution. As before, cell entries that deviate for at least 2% from the product of their latent marginals are printed in bold type. All of the bold entries are either cells representing total agreement or partial agreement.

Table 6.3.17

Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with two-variable effects as highest order interactions for interchangeable raters

			CONB					
			1		2		3	
NEUA=1	*NEUB*=1	*CONA*=1	**.03**	**(.00)**	**.02**	**(.00)**	.01	(.00)
		CONA=2	**.02**	**(.00)**	.01	(.00)	.01	(.00)
		CONA=3	.01	(.00)	.01	(.00)	**.03**	**(.00)**
	NEUB=2	*CONA*=1	.00	(.00)	.01	(.00)	.01	(.00)
		CONA=2	.00	(.00)	.01	(.00)	.01	(.00)
		CONA=3	.00	(.00)	.01	(.00)	**.04**	**(.00)**
	NEUB=3	*CONA*=1	.01	(.00)	.00	(.00)	.00	(.00)
		CONA=2	.01	(.00)	.01	(.00)	.01	(.00)
		CONA=3	.00	(.00)	.00	(.00)	**.02**	**(.00)**
NEUA=2	*NEUB*=1	*CONA*=1	.00	(.00)	.00	(.00)	.00	(.00)
		CONA=2	.01	(.00)	.01	(.00)	.01	(.00)
		CONA=3	.01	(.00)	.01	(.00)	**.04**	**(.00)**
	NEUB=2	*CONA*=1	.00	(.00)	.00	(.00)	.00	(.00)
		CONA=2	.00	(.00)	**.02**	**(.00)**	**.02**	**(.00)**
		CONA=3	.00	(.00)	**.02**	**(.00)**	**.10**	**(.01)**
	NEUB=3	*CONA*=1	.00	(.00)	.00	(.00)	.00	(.00)
		CONA=2	.01	(.00)	.01	(.00)	.01	(.00)
		CONA=3	.00	(.00)	.01	(.00)	**.04**	**(.00)**
NEUA=3	*NEUB*=1	*CONA*=1	.01	(.00)	.01	(.00)	.00	(.00)
		CONA=2	.00	(.00)	.01	(.00)	.00	(.00)
		CONA=3	.00	(.00)	.01	(.00)	**.02**	**(.00)**
	NEUB=2	*CONA*=1	.00	(.00)	.01	(.00)	.00	(.00)
		CONA=2	.00	(.00)	.01	(.00)	.01	(.00)
		CONA=3	.00	(.00)	.01	(.00)	**.04**	**(.00)**
	NEUB=3	*CONA*=1	.00	(.00)	.01	(.00)	.00	(.00)
		CONA=2	.01	(.00)	**.04**	**(.01)**	**.03**	**(.01)**
		CONA=3	.00	(.00)	**.03**	**(.01)**	**.08**	**(.01)**

Note. Entries in bold type depict expected proportions that deviate from the predictions based on the marginals by more than one decimal. Entries in parentheses represent the product of the latent marginals.

The two peer raters agree with respect to the two constructs (overall agreement) in 31% of the times. The total rate of agreement is slightly higher than for the self- and peer report. The overall agreement is mainly due to agreement on high scores of conscientiousness in combination with sensitive but stable or neurotic individuals (18% out of the 31% fall into these categories). Partial agreement on conscientiousness (28% in total) mainly occurs for highly conscientious individuals (20%). A slightly different picture can be found for neuroticism, the two peers agree with respect to the 1st category of neuroticism if at least one of the two chooses the 1st category of conscientiousness and the other maximally the 2nd category; or if they agree that the target individual is highly conscientious. Partial agreement with respect to sensitive but stable targets is only estimated to appear in 4% of all times and only for the category combination of moderately and highly conscientious ratings. Partial agreement is higher for neurotic individuals 8% of the times; it is mostly expected for the same cell combinations as mentioned above. In total, the interchangeable peers agree in 59% with respect to conscientiousness and in 51% of the times for neuroticism. It seems to be harder to agree on neuroticism than on conscientiousness. The analysis of the structurally different raters yielded comparable findings.

Additionally, the two constructs seem to be related since the cell entries are always highest for combinations of 2nd or 3rd categories of one construct with 2nd or 3rd categories of the other construct. The relatively high cell entries for the latent cell combinations [1 1 1 1], [1 1 1 2], and [1 1 2 1] also fit into this result.

As for the model for structurally different raters the complete quadrivariate table can be decomposed into its bivariate sub-tables in order to explain all associations. Tables 6.3.18 to 6.3.24 present the latent bivariate distributions as well as the category-specific agreement rates, distinguishability indices, and the method bias type II coefficients.

Inspecting the expected proportions for neuroticism depicted in Table 6.3.18 reveals that all cells on the main diagonal are more frequently expected than given their latent one-variable marginals. The category-specific agreement rates in Table 6.3.19 quantify the overrepresentations on the main diagonal. The combinations [1 1] and [3 3] are about 1.6 times more frequently expected than predicted based on the marginals and the combination [2 2] is around 1.4 times more frequently expected. $\kappa = .27$ indicates relatively low agreement between the interchangeable raters for neuroticism.

Table 6.3.18

Cross-classification of expected proportions for the latent variables representing neuroticism

	NEUB			
	1	2	3	
na = 1	.13 (.08)	.10 (.10)	.07 (.10)	.29
na = 2	.10 (.10)	.17 (.12)	.08 (.13)	.35
na = 3	.07 (.10)	.08 (.13)	.21 (.13)	.36
	.29	.35	.36	

Note. Values in parentheses represent the product of the latent marginals.

The disagreement cells besides the main diagonal reflect an interesting pattern of association. There is no reduction in disagreement compared to agreement on chance for the cell combinations [1 2] and [2 1]. That is, the peers do not distinguish well between the categories not neurotic and sensitive but stable. All other disagreement cells are less frequently expected than predicted by chance indicating that peers are able to distinguish between these categories (see Table 6.3.19). The distinguishability index shows considerably low values for the categories [2 3] and [1 3]. That is, peers can very well distinguish if a target is neurotic or not, the other peer does agree and not confound being neurotic with another category.

Table 6.3.19

Distinguishability index and category-specific agreement rates for neuroticism

	NEUB		
	1	2	3
na = 1	1.60	1.00	0.70
na = 2	1.00	1.42	0.62
na = 3	0.70	0.62	1.62

Table 6.3.20 presents the latent bivariate distribution for the two trait variables measuring conscientiousness. There is considerable agreement reflected in high proportions on the main diagonal. Peers *A* and *B* agree 3 times more frequently than predicted by the product of the marginals with respect to the first category of conscientiousness (category specific agreement rates on the main diagonal). They agree about 1.4 times more often than predicted by the marginals with respect to the 2nd and 3rd category of conscientiousness. Absolutely, the most entries can be found in the agreement cell for high conscientious target persons. 40% of all ratings fall into this category. Nevertheless, $\kappa = .32$ indicates relatively low overall agreement.

Table 6.3.20

Cross-classification of expected proportions for the latent variables representing conscientiousness

	CONB			
	1	2	3	
ca = 1	.06 (.02)	.07 (.05)	.03 (.08)	.15
ca = 2	.07 (.05)	.14 (.10)	.11 (.17)	.31
ca = 3	.03 (.08)	.11 (.17)	.40 (.29)	.54
	.15	.31	.54	

Note. Values in parentheses represent the product of the latent marginals.

Table 6.3.21

Distinguishability index and category-specific agreement rates for conscientiousness

	CONB		
	1	2	3
ca = 1	3.00	1.40	0.38
ca = 2	1.40	1.40	0.65
ca = 3	0.38	0.65	1.38

Table 6.3.21 depicts the distinguishability and category-specific agreement rates for conscientiousness. Peers confound the two lower categories of conscientiousness but can well distinguish between the highest and the other two categories of conscientiousness. Being traited seems to be a good moderator for agreement ratings of conscientiousness (see Funder, 1995).

Table 6.3.22 presents the latent joint distribution of the trait variables belonging to one rater (recall that the parameters are identical for the two raters). This cross-classification corresponds to the heterotrait-monomethod association sensu Campbell and Fiske (1959) for interchangeable raters. The absolutely highest deviations from the expected cell proportions from the product of the marginals can be found for the cell combinations [1 1], [1 3], and [2 3]. If one peer judges a target to be not neurotic the judgment will also more probably be not conscientiousness than predicted by chance and in the same vain more probably not highly conscientious. Sensitive but stable rated individuals will more probably also be rated to be conscientious than not conscientious. All other categories do not show strong deviations from the products of the latent marginals. These over- and underrepresentation in the joint distribution (combinations of categories) may be due to true associations between the latent constructs but may also be due to rater specific effects.

Table 6.3.22

Cross-classification of expected proportions for the latent variables originating in one peer rater

	CONA			
	1	2	3	
na = 1	.08 (.04)	.09 (.09)	.11 (.16)	.29
na = 2	.02 (.05)	.10 (.11)	.23 (.19)	.35
na = 3	.05 (.05)	.12 (.11)	.19 (.19)	.36
	.15	.31	.54	

Note. The model yields exactly the same results for the peer report *B*. Values in parentheses represent the product of the latent marginals.

Table 6.3.23 presents the latent bivariate distribution of neuroticism rated by one peer and conscientiousness rated by the other peer. This cross-classification corresponds to the heterotrait-heteromethod association sensu Campbell and Fiske (1959). The joint ratings of the different peers are not influenced by rater specific effects and, therefore, represent bias-free rates of over- or underrepresentation. This table shows that not conscientious individuals are also rated not to be neurotic. This may be due to a real association of the two categories but also due to ambiguous signals sent out by the target which may be interpreted as indicating not to be conscientious by one rater and not to be neurotic by the other. Additionally, the latent cell combinations of highly neurotic and moderately conscientious as well as sensitive but stable and highly conscientious are more frequently expected than predicted based on the product of the marginals.

Table 6.3.23

Cross-classification of expected proportions for the latent variables of different constructs originating in different peer reports

	CONB			
	1	2	3	
na = 1	.08 (.04)	.08 (.09)	.13 (.16)	.29
na = 2	.04 (.05)	.09 (.11)	.22 (.19)	.35
na = 3	.03 (.05)	.14 (.11)	.18 (.19)	.36
	.15	.31	.54	

Note. The model yields exactly the same results for the opposite combination (*NEUB* and *CONA*). Values in parentheses represent the product of the latent marginals.

Table 6.3.24

Method bias type II for peer reports

	CONA		
	1	2	3
na = 1	1.00	1.00	0.85
na = 2	0.50	1.11	1.05
na = 3	1.67	0.86	1.06

Note. The model yields exactly the same results for peer rater *B*.

Table 6.3.24 presents the method bias type II coefficients comparing the association within methods to the association across methods. The table reveals that monomethod associations differ from the heteromethod associations most strongly for the combinations of low conscientious targets which are not perceived as sensitive but stable (middle category) by one rater but much more often as neurotic. Single raters overestimate the association of highly neurotic and not conscientious as they also underestimate the association of not conscientious and sensitive but stable. All other associations do virtually not differ from the associations found between raters. One may speculate that targets who are rated as at least moderately conscientious are rated by their peers with a smaller bias. Conscientiousness may be seen as a visibility indicator for neuroticism.

6.3.2.2 Summary of the findings for the CT MTMR models for interchangeable raters

Although the CT-MTMR models for interchangeable raters are much more restricted than the models for structurally different raters and therefore should be more parsimonious, the applications for the CT MTMR model for interchangeable raters yielded the same computational difficulties that could also be found for the models with structurally different raters and, additionally, suffered from estimation problems concerning the equality restrictions in LEM (Vermunt, 1997a). The models with higher order interactions yield aberrant parameter estimates and can not be interpreted since they do not imply the correct structure for the expected proportions (the estimation yielded theoretically impossible results).

Yet, the model with two-variable effects as highest order interactions yields (relatively) sound parameter estimates which allow for analyzing the latent bivariate relationships. An inspection of the log-linear parameters is not meaningful with respect to this model because the only way to estimate the model in LEM is by means of dummy coded effects. Therefore, the parameters do not directly relate to the expected proportions and, moreover, cannot be interpreted in the ways described.

The analysis of the model with two-variable effects as highest order effects can be carried out with respect to the convergent and discriminant validity, method bias, and accuracy of the interchangeable raters. There is a considerable overall agreement rate showing that in about 1 out of 2 cases peer raters agree with respect to at least one of the two constructs. Inspecting the expected proportions may lead to the conclusion that agreement is highest for cell combinations of highly conscientious and sensitive but stable (middle category) as well as neurotic individuals. The partial agreement rates also show that peer raters agree more often for individuals classified in one of the above mentioned categories.

Since the CT MTMR model with two-variable interactions as highest order interactions is a hierarchical model it reproduces the latent bivariate joint distributions allowing for a direct interpretation of the expected bivariate proportions and the latent one-variable marginals. The category-specific agreement rates can be calculated to identify the overrepresentation in the cells on the (agreement) main diagonals. There is agreement for all cells on the two bivariate main diagonals (for neuroticism and conscientiousness). The category-specific agreement rates show that there is a much higher agreement for the combinations of the 1st categories for both traits. However, this does not imply that these rates are absolutely very high but relatively with respect to what can be expected knowing the marginals. κ is not very pronounced indicating low agreement rates. If there were a possibility to estimate the saturated model with correct standard errors one might inspect the corresponding log-linear effects to judge the ratio of agreement more exactly.

The distinguishability index reveals if disagreement cells are over- or underrepresented with respect to the product of the latent marginals. This index can be used to detect sources of disagreement. For conscientiousness this index revealed for example that peer raters confound the first two categories (lack of distinguishability). All other categories can be relatively well distinguished from each other for the two traits.

The cross-classification of trait variables belonging to the same method (heterotrait-monomethod associations) revealed that there are specific over- and underrepresentations of particular joint categories of neuroticism and conscientiousness. Being rated as not conscientious is related to being rated as not neurotic. Highly conscientious individuals are more probably rated to be sensitive but stable (middle category). This is true for one rater but also true for different raters. The method bias type II reveals differences with respect to the monorater bivariate joint distributions (categories) and the heterorater bivariate joint distributions for different traits. This index shows that considerable differences can only be found for individuals who are rated not to be conscientious.

In sum, I conclude that the CT MTMR model could be used to detect category-specific sources of convergence, category-specific lack of discriminant validity as well as distinguishability, it allows for a comparison of within raters associations across traits to estimate the rater-specific biases, and (theoretically) to examine if higher agreement rates are due to two-, three-, and / or four-variable effects. That is, if higher or lower degrees of convergent validity can be found depending on moderators of agreement (see Funder, 1995).

6.4 Discussion of the CT MTMR Models

In empirical applications, the CT MTMR models defined in this dissertation could provide a useful tool for the analysis of convergent and discriminant validity, rater bias, and determinants as well as moderators of agreement. However, to date these models cannot be soundly estimated prohibiting a proper interpretation of their log-linear parameters (estimation problems in LEM, Vermunt, 1997a). The practical applicability of the CT MTMR models depends on the availability of new software packages that overcome the current estimation problems.

Due to these estimation problems the expected proportions are reported and, more importantly, the empirical findings should not be substantively interpreted, instead the empirical applications serve to illustrate the newly developed models and the possibilities to interpret different model parameters. All different model parameters and their meaning for the analysis of the convergent and discriminant validity are explained in detail. Moreover, one can determine if one or more variables can be conceived to characterize a moderator of agreement (Funder, 1995). For instance, dependable individuals (as a concept related to conscientiousness) seem to be more congruently rated by peers (see Colvin, 1993b).

I define the method bias type I and II coefficient revealing information about different prevalence rates and different presumed associations of the rater. The distinguishability index provides information as to which categories can be neatly differentiated and which categories can be easily confounded. Additionally, the meaning of all different log-linear parameters with respect to the categories (agreement, partial agreement, simple agreement, and disagreement) they affect is exemplified and linked to the analysis of convergent and discriminant validity. These parameters provide pieces of information that cannot be retrieved from standard rater agreement models or latent rater agreement models.

7 Summary and Discussion

The aim of this dissertation was to develop i) latent rater agreement and ii) latent Multitrait-Multirater (MTMR) models for multiple categorical response variables in order to provide psychometric models for the analysis of convergent and discriminant validity corrected for influences due to measurement error. Additionally, indices have been defined that allow for the analysis of category-specific agreement rates, rater bias, and the distinguishability of the latent categories. Furthermore, the influence of particular latent statuses on agreement and / or disagreement may be analyzed. The focus was on the model development and the interpretation of the model parameters in terms of the analysis of convergent and discriminant validity, rater bias, and rater agreement (accuracy). Due to the computational difficulties and the logic of log-linear modeling with latent variables all definitions and applications are restricted to the case of maximally two traits judged by two raters.

This restriction is justified for several reasons. To date, the computational difficulties encountered during the estimation process prohibit the application of the models to more complex data sets. The definition of models for two raters comes closest to the inspection of the bivariate relationships in CFA MTMM models—the extension to the case of more than two raters and / or more than two traits is, in principle, straightforward for the CT MTMR models but adds higher order interactions to the model definition for each newly introduced latent variable. Therefore, it is impossible to give a global model definition; it is only possible to give model definitions relying on the number of traits and raters. The focus is on showing the strength of the newly developed models for two raters revealing pieces of information that are not available in standard rater agreement models. All models have been defined with respect to structurally different and interchangeable raters.

In a first step, I will summarize and discuss the newly introduced models with respect to the new pieces of information that can (theoretically) be gained. I will refer to the results of the empirical applications being aware that these may only be interpreted for illustrative reasons due to the estimation difficulties. In a second step, the models will be embedded in the larger context of research on rater accuracy (Funder, 1995) combined with the perspective of Multitrait-Multimethod measurement. Finally, the limitations of the newly developed models will be discussed and future research directions will be derived.

7.1 Summary of the Model Parameters and Model Results

7.1.1 Latent Rater Agreement Models

The first research goal was to develop latent rater agreement models which, in principle, are equivalent to Monotrait-Multirater models. Defining these models allows for an examination of agreement and disagreement of two raters on one particular construct. Raters can agree in a general way yielding constant agreement rates across all categories of the latent trait but raters may also agree in a more specific way showing high agreement rates for some latent categories and smaller agreement rates for other latent categories. Assumptions about constant agreement can be tested using the quasi-independence II models for structurally different and interchangeable raters but may also be tested restricting the log-linear two-variable effects for cells on the main diagonal to be constant in all other models (saturated, quasi-, and symmetry models).

Disagreement may also be modeled in the newly proposed models. Raters may well distinguish between particular latent categories from each other but have difficulties to distinguish between others. In these cases, disagreement rates for the first categories will be low and they will be higher for the “difficult” categories. High disagreement rates may point to categories that may be easily confounded by raters but that are theoretically distinct from each other or it may point to a lack of convergent validity within one trait. Examining a rating scale the models may be used to check if several categories may be empirically distinguished from each other. High expected proportions (or log-linear parameters around 1 or higher) for disagreement cells may show that the category definition should be optimized.

Disagreement is strongly related to rater bias. Rater bias may be analyzed yielding very detailed information about the categories that are more or less prone to the rater-specific effects. The first type of bias concerns different prevalence rates in the univariate latent distributions of the raters [method (rater) bias type I; see Agresti, 1992].

Additionally, the validity of the different items can be examined by an inspection of the effect-parameters or the conditional response probabilities. Strong effect-parameters indicate “marker” items (categories) for latent categories and may be statistically tested relying on the *z*-values of the underlying log-linear two-variable effects of the measurement model. However, these effects turned out to be very prone to boundary solutions and cannot soundly be interpreted in the presented applications. The conditional response probabilities however can be interpreted. Very clear implied typical response patterns indicate more reliable prototypical classifications and more easy to distinguish categories. The mean assignment probabilities (not available in LEM) indicate the reliability of the classifications.

The dimensionality of the response categories may also be examined using latent rater agreement models or log-linear models with one latent variable. Log-linear models with latent variables may be administered to ordered categorical ratings. If the categories follow the presumed ordered structure the model estimation will principally yield ordered latent categories with response categories reflecting a general increase or decline in their conditional response probabilities for increasing or declining response options (see e.g., Dillon & Mulani, 1984; Heinen, 1996, Langeheine, 1988).

The meaning of all model parameters will be summarized with respect to the saturated model because this model is most general containing all effects described below. All other rater agreement models are restricted versions of this model. See Figures 5.2 and 5.4 on how to obtain the more restricted rater agreement models. The meaning of the different variables and their effects or associations will be highlighted following a prototypical sequence of model inspection in empirical applications. As a prerequisite the model must show an adequate goodness-of-fit to the data in order to provide soundly interpretable model parameters.

Meaning of the latent variables and validity of their indicators. The meaning of the latent variables in log-linear models with latent variables can principally not be known beforehand but has to be determined inspecting the empirical results (except for the case of a priori restricted model parameters). The direction and the strength of the link between the latent variable and its indicators determine its meaning in models with unordered categorical latent variables (see e.g., Hagenaars, 1993). This examination is nothing else but the analysis of validity of the latent variable or its measures (Messick, 1989). In general, different statuses on the latent variable must produce different expected scores on its manifest indicators. The two latent variables in latent rater agreement models must at least approximately represent identical categories in order to allow for an examination of rater agreement. This is a crucial point for the analysis of rater agreement (and also for CT MTMR models). In the best case, the conditional response probabilities do not differ across raters. If the conditional response probabilities are not identical theoretical considerations about and interpretation of the conditional response probabilities may still guarantee that the two judges rate the same construct (see Section 4 for more details).

One may examine the reliability and / or validity in models with categorical latent variables. The conditional response probabilities indicate the degree to which a given category of an indicator can be conceived as a good representation of the latent category. In the same vain, effect-parameters or odds can be used to examine the convergent validity of indicators in a way closely related to the inspection of the conditional response probabilities. One may conclude that an indicator is a valid (good or marker) indicator of a latent category if it shows strong effect-parameters (see Section 4). However, there are no benchmarks as to which size of a conditional response probability or effect-parameters may be considered showing a strong measurement relationship. The interpretation of these parameters depends on the research domain and prior results.

There is not *one* parameter representing the relation between the manifest variable and the latent variable in models with categorical latent variables but there are as many parameters as there are combinations of latent and manifest categories. That is, there are nine log-linear two-variable effects indicating the interplay between a three-categorical latent and a three-categorical manifest variable. This allows for a detailed analysis of validity with respect to the categories.

Consider neuroticism in the self-report. There are three latent categories which can be clearly distinguished with respect to their conditional response probabilities. The categories can be interpreted as three types of neuroticism (not neurotic, sensitive but stable, and neurotic) in a theoretically meaningful way inspecting the nine conditional response probabilities. The conditional response probabilities change quite heterogeneously across the different items and latent categories. Different items may be used to distinguish between the different latent categories. The neurotic personality type can be easily separated from the sensitive but stable and the neurotic personality type inspecting the typical response behavior for items “vulnerable” and “sensitive” (large differences in the conditional response probabilities). That is, these two items validly separate the first latent category from the other two latent categories. Yet, these two items cannot be used to separate the sensitive but stable from the neurotic personality type. The typical responses for these two latent classes are responses using the neurotic category. These items thus do not validly separate the middle and the highest latent classes from each other. In the same vain, items “moody” and “self-doubtful” can be regarded as valid indicators to separate not neurotic as well as sensitive but stable individuals from neurotic individuals because the conditional response probabilities differ to a great extent. These two items do not very well discriminate between the latent categories “not neurotic” (lowest category) and “sensitive but stable” (middle category).

On the one hand, this examination of validity is complex in the models with categorical data; on the other hand, this examination allows for a better understanding of the latent categories and the associations between the manifest response variables. The conditional response probabilities imply that sensitivity and vulnerability are easier to feel or perceive than moodiness and self-doubtfulness. However, this assumption should be examined in detail additionally relying on models of Item-Response-Theory (IRT) as the graded-response model (Samejima, 1969), for example. I also must emphasize that I collapsed two times two response categories to avoid computational difficulties. This certainly also afflicts the interpretation of the model results with respect to the difficulties of the items and the ordered structure of the latent and manifest categories.

Latent one-variable distributions. Inspecting the latent one-variable distributions reveals if the two raters perceive the same prevalence rates for the construct under consideration. In general, their prevalence rates should not differ to a large extent from each other to still reflect the same construct (Zwick, 1988). If the prevalence rates differ considerably the raters judge different phenomena. However, there are no guidelines as to which differences in the prevalence rates can be considered meaningful. This problem is not examined in this contribution. The difference of the latent distributions is quantified by the rater bias type I coefficient. This coefficient compares the expected proportions of identical categories of different ratings. One may test if the latent marginals are homogenous (identical) comparing models with and without equality restrictions on the latent univariate distributions. However, this is not in the focus and thus was not done. The focus is on the possible interpretations of the latent one-variable parameters and the pieces of information they provide on agreement and disagreement as well as rater bias. For the rater agreement models as well as for the CT MTMR models, all rater bias I coefficients fell into the range of $MB1_{(ns=1.na=1)} = 0.83$ to $MB1_{(ns=3.na=3)} = 1.15$ for neuroticism and $MB1_{(cs=1.ca=1)} = 0.79$ to $MB1_{(cs=3.ca=3)} = 1.41$ for conscientiousness (comparing the self-report with the peer report). The relatively high rater bias I coefficient for the 3rd category of conscientiousness has been interpreted in Section 6.3. To my mind, it is plausible to assume a peer bias towards higher conscientiousness ratings due to the composition of the sample and some aspects of conscientiousness that may not be openly displayed (e.g., fighting against a tendency of procrastination).

The rater bias type I reveals if a category for one rater is strongly overrepresented with respect to the identical category for the other rater and, additionally, if the necessary underrepresentation of the other categories is found for one or for several categories. The rater bias type I coefficients may reveal that one rater shows more ratings in the highest category than the other rater but that both show equally frequent ratings in the lowest category. Therefore, one may conclude for ordered categories that the first rater has a lower “threshold” to pass from the middle to the highest rating categories.

Latent two-variable distribution. The latent two-variable distribution reveals to which extent the two raters agree or disagree. κ may serve as an indicator of overall agreement. High agreement indicates convergent validity on the level of trait variables. Additionally, category-specific agreement rates may be calculated indicating for which categories agreement can be found. Without further analyses on the decision making process, one may conceive these categories as *good categories* upon which raters easily agree. In general, some categories of a categorical trait are much easier to agree upon than others (see e.g., the concept of visibility, Funder, 1995). The latent rater agreement models allow for determining good categories inspecting the category-specific agreement rates and the two-variable log-linear parameters for agreement cells. High and significant parameters indicate good categories. Validity determined as agreement concerns absolute agreement. Any slightest form of disagreement (i.e., one rater choosing the risk-seeking category and the other rater the gambling category) is related to a decrease in convergent validity (although disagreement with respect to gambling and risk-seeking might be less striking than disagreement between gambling and security oriented). This problem may be circumvented by accepting disagreement for closely related categories as still refelcing vonvergent ratings or by collapsing these categories.

In applications showing agreement, disagreement cells will be less frequently observed than predicted by the product of their latent marginal distributions. However, it is important to know which categories can be neatly distinguished from each other and which may still be confounded. I propose the category distinguishability index to examine the ratio to which the expected proportion of the disagreement cell deviates from the product of the marginals. Very small values (close to 0) indicate that raters can very well distinguish between the corresponding categories. Values close to 1 indicate that the association of the latent categories corresponds to the association one would expect for independent categories. In this case, raters do not confound the categories (this would be indicated by a distinguishability index larger than 1) but they also do not distinguish well between these categories. Therefore, values close to 1 or above 1 indicate the need to clarify the category definitions of the items or to train raters. Additionally, the distinguishability index gives some hints on effects of possible moderators of accuracy (agreement). If for a specific category of one trait there are very low distinguishability indices and there is very high category-specific agreement, the specific category is a good category.

Special rater agreement models. Specific patterns of agreement and disagreement can be modeled adopting the rater agreement models for observed variables to log-linear models with latent variables. The latent saturated model does not impose any restrictions on the associations between latent categories.

If the patterns of disagreement (the distinguishability indices) can be approximately mirrored on the main diagonal the quasi-symmetry model may be a good representation of the data. This model implies that the two raters distinguish the different categories in a similar way. If additionally their latent marginals are homogenous they distinguish categories in perfectly the same way (symmetry). For example, the two raters confound the categories of being neurotic and moderately conscientiousness to the same ratio as the inversed combination being moderately conscientiousness and neurotic (keeping the ordering of the raters the same).

If raters distinguish equally well between all categories besides the main diagonal one of the quasi-independence models may fit to the data. In these models, the disagreement rates are constantly reduced since there are no associations besides the main diagonal. All associations in this model are due to chance (independence assumption) except for higher agreement rates on the main diagonal. Agreement can be overrepresented changing from category to category yielding quasi-independence I (reflecting more or less good categories) or be constant across categories yielding quasi-independence II (reflecting good judges and / or good traits). All rater agreement models can be derived from the saturated model implying meaningful restrictions (see Figures 5.2 or 5.4).

Structurally different vs. interchangeable raters. All models can be defined for structurally different and interchangeable raters. The summary of the meaning of the latent variables and the model parameters presented above is valid for the more general case of structurally different raters. The models for interchangeable raters differ with one major aspect from the models for structurally different raters. If raters are interchangeable they originate in the same population leading to the restriction of measurement invariance. Additionally, their latent distributions must be identical implying that they perceive the same prevalence rates and that they confound categories in identical ways. Therefore, only three rater agreement models exist for the case of interchangeable methods: The symmetry as well as the quasi-independence I and II models. The inspection of rater bias type I is meaningless in these models by definition; all other indices can be interpreted as presented above.

7.1.2 Multitrait-Multirater (MTMR) Models

The second major research goal was to extend the latent rater agreement models to allow for the analysis of more than one trait. Most emphasis was paid to the interpretation of the different log-linear parameters in the saturated CT MTMR model allowing for a detailed analysis of agreement and disagreement (reflecting convergent and discriminant validity as well as method bias). In this model, moderators of agreement and disagreement can be identified relating the MTMR model to the realistic accuracy model (RAM, Funder, 1995). Having estimated a CT MTMR model for categorical data the following steps should be taken to investigate the results:

First of all, only models showing an adequate fit to the data should be examined. In fitting models the *meaning of the latent variables and validity of their indicators* have to be analyzed. The inspection of the effect-parameters or the conditional response probabilities has to be executed as described above. The distributions of the latent marginals can be inspected to identify if the prevalence rates are the same for identical constructs across raters. Method (rater) bias type I can be determined as described above. The analysis of agreement (convergent validity), disagreement, bias of ratings, discriminant validity and moderators of agreement and disagreement is rather complex in CT MTMR models.

Agreement. As a great advantage the saturated CT MTMR model simultaneously allows for an inspection of agreement and moderators of agreement. Agreement between raters may differ across categories of the latent traits under consideration. Additionally, the CT MTMR model allows for determining if agreement differs with respect to ratings on the second construct. The model reveals if there is higher agreement on neuroticism for highly (congruently rated) conscientious individuals, for example. I will shortly repeat the theoretical impact of the log-linear effects of different levels on agreement and convergent validity (see Section 6 for more detailed explanations). The empirical application revealed that the models with higher order effects could not be soundly estimated.

Conditional complete agreement is depicted by the four-variable log-linear effects for cells indicating simultaneous agreement on the two constructs. The log-linear parameters indicate the odds for complete agreement to the expected agreement due to all lower order effects. They, therefore, reflect above chance complete agreement where chance complete agreement is the expected agreement given all lower order effects. Complete agreement on the two constructs could also be produced by the one- and two-variable effects but not by the three-variable effects. Conditional complete agreement may be constant for all cells or category specific (see Section 6 for a thorough discussion). Constant complete agreement is related to a property of targets as being good targets. If raters agree on one target's first trait (conscientiousness), they also agree on this target's second trait (neuroticism). Category-specific complete agreement is related to palpability (Funder, 1995). Palpability reflects the fact that some traits of some targets may well be identified whereas the same traits cannot be accurately judged for other targets. In the CT MTMR model, palpability is more fine-graded as it may also occur that some targets may only be better judged for given combinations of categories. The heuristic inspection of the latent quadrivariate distribution revealed that highly neurotic individuals are more easily congruently judged being highly conscientious. Inspecting the four-variable log-linear effect for this cell combination [3 3 3 3] would reveal if this complete agreement was due to lower order effects or due to the palpability effect.

Partial agreement depends on log-linear effects on different levels. Four-variable effects depict if specific constellations of disagreement on one construct co-occur more frequently with agreement on the second construct. These effects indicate differential views of the raters about the association of the two constructs. Agreement on high conscientiousness may be associated to different ratings on neuroticism (e.g., neurotic by the self-report and sensitive but stable by peer report *A*). If one of the raters is outstanding and may provide a better approximation of the true status, the interpretation of the three-variable effects as influencing the partial agreement becomes meaningful. If, for example, the self-raters judge themselves as not conscientious it may be the case that self- and peer raters more easily agree on the not neurotic category. This may be due to the fact that being not conscientious is a moderator of agreement (visibility indicator) for low neuroticism.

"Simple agreement". Agreement may also be analyzed at the level of bivariate relationships. If one is interested in agreement rates without any further information about the genesis of agreement the log-linear parameters (or cells of the bivariate distribution) representing agreement can be examined as described above. However, these parameters represent main effects which may change with respect to different constellations on the other variables.

Disagreement. The CT MTMR model is suited for the analysis of the genesis of disagreement. In principle, disagreement should be expected to a lower extent than predicted by the product of the latent marginals. Yet, the distinguishability index (the ratio of the expected proportion to the product of the latent marginals) will differ across category combinations indicating that raters can very well or less well distinguish between pairs of categories. This distinguishability can be stable across all category constellations on the other trait but also differ with respect to the other trait.

The four-, three-, and two-variable effects reveal if specific patterns of disagreement are more or less often expected than other patterns. *Partial agreement* as described above is a special case of disagreement since the two raters agree on the other construct. If there is high partial agreement an analysis of the decision making process for the construct upon which the two raters disagree may be worthwhile to reveal if the same behavioral cues are perceived and if they are interpreted in the same way.

Disagreement which is due to four-variable effects shows to which degree the two raters weigh information differentially. There is, for example, some confusion about targets who are rated sensitive but stable in the self-rating and neurotic by peer *A* with respect to the ratings of moderate and high conscientiousness (see Table 6.3.2). Peers have a high probability to rate targets higher on conscientiousness if targets perceive themselves as moderately conscientious.

Three-variable effects depict if there are specific categories of one construct that are associated to high rates of disagreement. Reconsider the example of a target person who does not show her or his feelings. This individual will hardly be congruently judged by peer raters concerning the momentary emotional status. The three-variable parameters may thus indicate categories being moderators of disagreement.

Disagreement mirrored by the two-variable log-linear effects shows the principle disagreement “averaging” across all higher order moderator effects. Inspecting the two-variable effects reveals which categories cannot be well distinguished by the two raters. If there is no higher order effect, the analysis of the decision making process concerning categories that are too easily confounded may help to improve rater agreement and reduce disagreement. However, if higher-order effects are present it is these effects that indicate under which conditions peers agree and disagree. Knowing these specific constellations allows for a more precise and stringent analysis of the decision making process.

Rater bias. Rater bias can be analyzed relying on different indices. The method (rater) bias type I coefficient reveals if there are differences in the latent marginal distributions for the same construct. This bias should not be very pronounced allowing for an investigation of rater agreement. If raters differ extremely in the prevalence rates of their ratings the examination of rater agreement becomes meaningless (Zwick, 1988).

Raters may also show biased ratings with respect to the categories of different constructs they associate. The rater bias type II index compares the associations between different categories across constructs of one rater to the expected association across raters. This index reveals if raters have a specific view as to which categories of the latent traits are more or less associated than expected for different raters. The definition of this coefficient as the ratio of a multitrait-monomethod to the multitrait-heteromethod associations is related to the logic of direct product models (see e.g., Browne, 1984, Oort, 1999; Wothke & Browne, 1990).

However, if higher order effects are present, the rater bias type I and II coefficients can only be interpreted as average effects which may be moderated as are two-variable effects in models with higher order interactions. The CT MTMR model principally allows for determining specific category constellations which may lead to especially biased ratings or to a reduction in bias. However, in the current applications these constellations could not be identified relying on the log-linear parameters since the model estimation yielded almost no higher order effect without boundary solution. Yet, inspecting the expected proportions of Table 6.3.2 in a heuristic way may lead to the hypothesis that peer *A* associates moderate conscientiousness to being neurotic and high conscientiousness to being sensitive but stable for targets who rate themselves being sensitive but stable and moderately or highly conscientious. It would be very interesting to analyze if these effects are due to three-variable interactions or four-variable interactions in soundly estimated models. If these effects are not due to four-variable interactions but to three-variable interactions the self-ratings on one construct influence the joint ratings of the peer raters. This influence is then independent from the self-rated score on the other construct. However, there may also be an effect of this other self-rated trait on the joint peer ratings. This could be interpreted as two different and independent effects representing the peer-specific view as a function of the different statuses in the self-report. If these effects are additionally due to a four-variable interaction a “joint halo-effect” may be present implying that one constellation in the self-ratings produces a particular joint bias in the peer ratings. Detailed inspections of the answer process may help to enhance rater agreement and reduce rater bias.

Discriminant validity. The discriminant validity can be analyzed relying on different associations. The simplest case in the model with two-variable effects as highest order interactions has been illustrated in detail. The inspection of the latent bivariate associations is closely related to the inspection of heterotrait-heteromethod and heterotrait-monomethod associations as described by Campbell and Fiske (1959). The application for structurally different raters revealed that the cross-classification of neuroticism in the self-report with peer-reported conscientiousness showed almost perfect discriminant validity. This was also true for most of the cells in the heterotrait-monomethod cross-classification except for the combination of being not neurotic and not conscientious in the peer rating. This category combination is expected almost two times more often than expected by the latent marginals. The opposite is true for the combination of not neurotic and conscientious which is expected only half as often as predicted by the product of the marginals. If information on the log-linear parameters and their standard errors was available for the models with higher order effects the model parameters could indicate if discriminant validity remains stable across different constellations on the other trait or if it changes.

Determinants and moderators of agreement and disagreement. In principle the CT MTMR models allow for the examination of moderators of agreement and disagreement via their latent three- and four-variable log-linear effects. These moderators have already been discussed with respect to agreement and disagreement (see above). I will shortly repeat the possible moderators that can principally be detected in the CT MTMR models. Good categories may be identified as categories with very high agreement rates across raters. Good targets are targets upon whom raters agree on all constructs (i.e., especially consistent individuals, see Funder, 1995), this may also be the case for special combinations of good categories and good targets—this combination has been introduced as palpability for the interaction of traits and targets by Funder and extended to the interaction of categories and targets in this dissertation. Good judges agree with each other independently of the category combinations. The CT MTMR model is restricted in its information about all possible determinants and moderators of agreement and disagreement because information is only available for two traits times two rater. It is not possible to separate some of the different moderators from each other to identify the different influences (see below). Additional information gained by more traits and raters must be used to get more insight into the moderating effects.

Structurally different vs. interchangeable raters. The CT MTMR models can be used to analyze structurally different and interchangeable raters. Interchangeable raters require special restrictions on the model parameters representing their interchangeability. These restrictions have been introduced in detail. All variables, effects, and parameters can be interpreted as for the case of structurally different raters.

7.2 The “Joint Framework” of the CT MTMR Model and the Realistic Accuracy Model

Validity and reliability are of highest importance in many areas in psychology. It is important that psychologists detect and correctly use the behavioral cues that indicate specific patterns of behaviors or latent typological variables (such as clinical disorders). The detection of the relevant cues and the processing of information executed by a psychologist, for example, can be described by models of signal detection theory (SDT; see Wickens, 2002). These models link the perceivable cues (visual, auditory, haptic, and olfactory) to a then activated category and to the mental registration. Analyzing these processes may be very helpful to explain how judges make up their minds depending on the cues they can perceive or the cues they even did not perceive concerning several items as “being moody”, “self-doubtful”, “sensitive” or “vulnerable”.

Funder (1995) introduced the realistic accuracy model (RAM) as a logic chain of determinants of accurate judgment. The knowledge about the properties of *good judges*, *good indicators*, *good targets*, and *good traits* as factors enhancing rater agreement may help to improve the quality of ratings or may help to explain why some ratings are inaccurate. Funder (1995) developed the RAM focusing on factors and their interactions that may enhance “rating accuracy”. He argues from a postpositivist perspective saying that “truth indeed exists but there is no sure pathway to it (p. 656)” relying on philosophical positions such as critical realism and pancritical rationalism (for more details see Funder, 1995). His approach is closely related to the approach of Brunswick (1956, cited after Funder, 1995). In Funder's point of view, accuracy (approximating truth) is enhanced when raters agree. Without engaging in a discussion about the existence of truth and the possibility to perceive or know it, his arguments seem to be true in the context of rater agreement, too, as his considerations and implications directly apply to the models of rater agreement. Thus they may help to get a deeper understanding of agreement and disagreement of multiple raters.

Agreement as a joint product of the target and the judge depends on four principal sources of agreement: i) the relevance of behavioral cues to a personality trait, ii) the extent to which these cues are available to observation, iii) the extent to which these cues are detected, and iv) the way in which these cues are used (Funder, 1995, p. 658). These four determinants are connected using a logical chain leading from the trait (construct) to the final ratings. i) A trait generally produces a behavioral effect which is conceived relevant for this trait; however, ii) this behavioral effect must be available to the judge to become meaningful with respect to the rating. Changes in cortical activity may generally not easily be observed whereas facial expressions such as flushing or smiling are. Additionally, iii) the judge must be attentive and able to detect these behavioral cues. The detection may be hampered by many factors such as inattentive judges, distracting situations, or situational factors which render a behavioral cue difficult to be seen—targets may look into another direction due to experimental instructions and therefore their flushing is difficult to be seen. Finally, iv) the judge must correctly link the behavioral cue to the trait it represents. A judge may believe a behavior to be diagnostic of a particular trait while it is diagnostic of another trait or of nothing at all.

In RAM four theoretically possible moderators of accuracy are introduced: i) good judge, ii) good target, iii) good trait, and iv) good information:

i) Good and bad judges can be differentiated by their abilities to detect and use readily available behavioral cues. Funder (1995) introduces three components rendering a judge a good judge. a) Experience and / or knowledge about personality traits and how they are revealed in behavior. b) General abilities such as intelligence or more narrow abilities as cognitive and attributional complexity may improve the possibility that detected cues are used in a valid manner. This corresponds to the analysis of information processing and can best be done using techniques of signal detection theory (see e.g., Wickens, 2002). c) Finally, motivational aspects may lead to more accurate judgments if the motivation to provide valid ratings is high (Flink & Park, 1991) but may also lead to a distortion of ratings. A person who has a strong need to be always in the right may not be a good judge judging a target's actions which are opposed to the rater's own beliefs (see Funder, 1995).

ii) Good targets can be judged correctly having relatively few information about their behavior (see e.g., Colvin 1993a, 1993b). Good targets are characterized by showing a high cue availability and relevance. Funder (1995) lists a number of hypotheses which might explain why some individuals are much easier to judge than others. Individuals with high activity levels should show more behaviors and, therefore, be more easily judged correctly. The same is true for talkative people. High self-monitoring individuals are more difficult to judge according to RAM because these individuals change their behaviors as a function of their surroundings. This is related to the question if individuals are traited (i.e., having a trait) or not. Baumeister and Tice (1988) introduced the concept of "metatraits" describing the phenomenon that some individuals act consistently over situations and, thus, have specific traits while others do not. Colvin (1993a, 1993b) found that some individuals are both more consistent in their behaviors and more likely to be agreed about as a result of their consistent behavior.

Conversely, "bad targets" are either individuals who are inconsistent in their behaviors (e.g., high self-monitoring) or individuals who conceal certain aspects of their behavior. Criminals for example may not overtly show criminal acts leading to agreement about their “non-criminality”, however these ratings are by far not accurate.

iii) Good traits are characterized by easily available and highly relevant behavioral cues. These are traits which are associated to easily observable behaviors such as positive social interaction for sociability and which are frequently displayed (e.g., a person who often seeks social interactions). In short, some traits are more *visible* than others. Visibility is closely associated with interjudge agreement (see e.g., Funder & Dobroth, 1987). However, to my mind visibility is not the same as availability and relevance. Some behaviors may be frequently available but relevant for different traits. Talkative individuals may be nervous, sociable, dominant, and / or all of the three. Therefore the behavioral cue "talkative" is easily available but ambiguous with respect to several traits. *Visibility* in my understanding is the interplay of availability and relevance of several combined behavioral cues being highly indicative for a particular trait. Visibility may also be enhanced by other properties of the individual making it easier to differentiate between different traits.

iv) Good information is the signals sent out by the target which might principally lead to accurate judgment. This moderator only concerns the availability of relevant information and not if raters perceive this information or the way they process the information.

These four moderators may each have an isolated effect on the accuracy of ratings but they may also interact. Traditional rater agreement models deal with one trait and multiple raters allowing for an examination of agreement and disagreement. Moderators at the level of targets and / or raters may be integrated in order to explain why there is agreement on some targets and why some raters agree while others do not. The interaction between these two moderators is called *relationship* (Funder, 1995). In RAM, *expertise* denotes the interaction between raters and traits. Expertise is high if a particular rater has enough knowledge about a given trait and its behavioral cues. *Sensitivity* characterizes the fact that some judges may be better in perceiving particular relevant information than others; however, this effect changes as a function of the kind of information and the rater. *Diagnosticity* names the fact that some traits can only be judged based on particular information and that the accuracy depends on the level of generalization of the trait (see Funder, 1995). *Divulgence* denotes the fact that some information about a target may help a rater to judge this target accurately while the same information concerning another target may not at all help to improve the rating quality. Individuals of different ethnic groups may show the same behavior but this information may not indicate the same concepts (shaking ones head is associated to saying no in western cultures but means yes in large parts of India).

All these interactions could be analyzed in rater agreement models if additional information was incorporated into the model. Yet, there is one interaction that—at least in parts—can be examined relating rater agreement models to each other as is done in the Multitrait-Multirater models: *Palpability* denotes the interaction of traits and targets. That is, certain traits might be easy to judge in some targets but not in others. Integrating additional information into rater agreement models such as multiple traits allows for a deeper understanding of which personality types may be congruently or more accurately rated. To my mind the interaction of targets and traits (palpability) is related to the visibility of a trait. That is, a highly extraverted individual may be much easier judgable on certain traits because she or he provides much more behavioral cues and is much more open-hearted. Therefore, being extraverted may also be conceived as an indicator of visibility for some traits.

Funder (1995) explicitly claims to enlarge RAM by integrating a multiple cues and multiple traits perspective. Without explicitly mentioning, Funder implies that Multitrait-Multirater models can be seen as a special case of multitrait RAMs. His approach provides some interesting theoretical considerations abut the different effects of the Multitrait-Multirater models for categorical data.

It is thus logic to combine the strength of the different approaches with each other yielding a “joint framework” for the analysis of rater agreement and disagreement. For example the “Children's Depression Rating Scale - Revised” (Poznanski & Mokros, 2004) is used to rate a child's status on a more or less abstract construct as depression based on a semi-structured interview. The categorization of a child as suffering from a depression depends on the classification as being sad, having morbid thoughts, failing at school and other classifications. This approach is psychometrically mirrored by the log-linear model with one latent variable.

In this example, it is of utmost importance that the ratings of the clinical psychologist are accurate for each child she or he has to rate. Therefore, new clinicians (trainees) should be trained to provide accurate ratings. The latent rater agreement models could be used to examine if the ratings of the trainee come close to what an expert rates (considering the expert as gold standard in the models for structurally different methods) or to compare the ratings of different trainees in order to inspect which category definitions should be made much clearer (using models for interchangeable raters).

Furthermore, it is very important to consider additional clinical symptoms to inspect if the trainees correctly detect comorbidity or if they have special beliefs or theories about which symptoms are related or not (rater bias). To this end the CT MTMR model could be used. It is also highly important to know more about the moderators of agreement (accuracy). Which forms of depression can be considered palpable? Are there some easily detectable forms of depression? Is agreement higher for particular children? What are these children's' personality traits or clinical disorders that render these children more judgeable? Do trainees distinguish equally well between all categories? Which categories do they confound more often than others? All these pieces of information could be detected by sound applications of the CT MTMR model extended to more than two traits and more than two methods. However, these are the typical questions of the rater accuracy model presented by Funder (1995). In order to differentiate between good judges and good targets, for example, more than two raters are needed. If there is a group of easily judgeable targets all raters will agree with respect to their latent statuses. If agreement is due to the good judges not all of the three or more judges will agree with respect to the targets. Additionally, more than two traits are needed to identify if there are good traits, this is the case if some traits (extraversion and sociability, for example) can be easily judged (visible or good traits) by all raters whereas others (e.g., neuroticism) are harder to judge. The same is possible for specific categories of particular traits and their combinations. Detailed analyses of the log-linear model parameters would enable researchers to identify these moderators of agreement (accuracy).

Although the CT MTMR model allows for a deeper understanding about determinants and moderators of agreement it is by no means a process oriented model. That is, to study the underlying process of decision making the implications of the realistic accuracy model (Funder, 1995) should be related to models of signal detection theory (SDT; e.g., Wickens, 2002). These approaches could be used to clarify which kind of information is good information and what availability means in the perspective of a cognitive psychologist.

Crucial questions in this field are: What behavioral cues have to be emitted to render a behavior judgeable? What kind of information leads to a valid judgment on item contents such as “vulnerable”? Is it best to perceive verbal, behavioral, and auditory cues simultaneously or in sequence? SDT may help to analyze these research questions at the beginning of the rater accuracy model (relevance and availability as well as perception of cues). The CT MTMR model (but also models of SDT) may rather be used at the end of the logic chain when several raters may be compared with respect to how they used the perceived information. Therefore, I consider the CT MTMR model as a model that may broaden the perspective of the rater accuracy model but also as a model that is—directly implied by the logic of the realistic accuracy model—to be at the end of the logical chain analyzing the ratings.

7.3 Limitations of the Models and Future Research Directions

The major limitation of the presented models is their computational complexity. To date no software package allows for a sound estimation of the log-linear parameters of the most complex MTMR models. Future research directions concern the development of better estimation procedures for the log-linear models with latent variables (Latent Class Models). If these are available the applicability of the CT MTMR model might be examined in simulation studies relying on empirical and / or simulated data sets. However, there may only be guidelines concerning sample size requirements because the identification of a model is always in parts an empirical issue (see Section 4.1.2). In the current application, no information could be gained if the complex MTMR models could not be estimated due to intrinsic non-identification or to non-identification due to a sparse data problem. Software packages that could be used to analyze the MTMR models should integrate several components: i) better estimation procedures as Bayesian estimation methods using prior information (Maris, 1999; Vermunt & Magidson, 2002, 2005), ii) an automatic identification check as implemented in PANMARK, for example (van de Pol, Langeheine, & de Jong, 1996), and iii) the possibility to run bootstrap analysis.

The MTMR model is not soundly applicable to the data set presented in Section 4. If researchers are confronted with similar problems as encountered in the empirical applications at least different latent rater agreement and Correlated Traits models can be estimated crossing all traits and methods. These models reveal information that may answer some of the research questions listed above. However, conditional effects (as three- or four-variable effects in the MTMR models) cannot be analyzed.

Future research should be conducted on analyzing large data sets which may be found in organizational psychology where many clients rate many employees. Consider a call-center where clients are oftentimes asked to rate some properties of the agent. A fixed number of clients could be randomly drawn for each agent and their agreement and disagreement as well as the convergent and discriminant validity of the evaluation scale could be analyzed. The more complex situation with differing numbers of clients for the agents could be solved adopting the multilevel-latent class approach introduced by Vermunt (2003, 2005, 2008).

Young physicians could be trained relying on the latent rater agreement models or on the MTMR models if they were asked to rate patients during the ward rounds. Their ratings could be compared with ratings of other young physicians on the same patients or with the ratings of the physician in charge. This information could be used to develop specific programs to train the accuracy of the young physicians.

Data sets containing missingness on observed data which are likely to occur in many applications will additionally increase the complexity of the estimation process. Vermunt (1996) proposed an approach to analyze models with unobserved (latent), partly observed, and observed data. In this approach, response indicators have to be used indicating the missingness. This results in additional model complexity and has not yet been defined for the models presented here.

If the CT MTMR model should proof to be applicable to empirical data situations the newly defined indices (method bias type II and distinguishability index) should be investigated in more detail. It should be examined if there is any meaningful benchmark or threshold as to which ratio is of substantial interest for given research domains. In settings with many raters, careful considerations as to which raters may provide bias free associations are necessary and will afflict the definition of this index.

The quasi-independence models offer interesting possibilities to model disagreement and agreement. If the CT MTMR model may be soundly estimated it may also become meaningful to investigate the structure of agreement restricting the monotrait-heteromethod associations in a larger model. In the same vain, the structures of rater agreement might be adapted to three- and four-variable effects yielding non-hierarchical higher-order rater agreement models. In these models, all effects may be removed from the saturated model that do not relate to a simple, partial, or conditional overall agreement. This model would imply random associations for complete disagreement cells and might give important insights into rater bias. The clear psychometric definition and the interpretation of the log-linear parameters will be tedious because these kinds of models are no longer hierarchical. However, it might be adequate for rather distinct raters who might be expected to agree more often on some of the constructs but not to show related joint ratings for other constructs. Imagine colleagues and supervisors as raters of a target working in the service sector where workers are asked to be especially friendly and helpful. In this case, a 3rd order quasi-independence model could be fit allowing only for higher overall agreement rates. The target person will most probably be rather friendly and helpful to all clients and especially friendly and helpful if the supervisor is present (concealing some of her or his traits), but she or he may also be much less friendly and helpful with some of her or his colleagues. Therefore, the ratings of the supervisor and the colleagues will most probably differ from each other.

The latent rater agreement and the CT MTMR models presented in this dissertation may reveal important information about the convergent and discriminant validity of ratings in empirical applications. To date, only the less complicated models can be soundly estimated. If there are more advanced and more efficient estimation procedures the CT MTMR model may become applicable and its strength and gain in information concerning ratings can be used to enhance the quality of psychological ratings and to understand more about the determinants and moderators of agreement and disagreement.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Agresti, A. (1992). Modeling patterns of agreement and disagreement. *Statistical Methods in Medical Research, 1*, 201-218.

Akaike, H. (1974). A new look at statistical model identification. *IEEE transactions on Automatic control*, *19*, 716-723.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*, 317-332.

American Psychiatric Association (2000). D*iagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision* (*DSM-IV-TR*). Washington, D.C.: American Psychiatric Association.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 42*, 69-82.

Bakeman, R., & Gnisci, G. (2006). Sequential observational methods. In: M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 127-140). Washington, DC: American Psychological Association.

Baumeister, R. E., & Tice, D. M. (1988). Metatraits. *Journal of Personality*, *30*, 571-597.

Biesanz, J. C., & West, S. G. (2004). Towards understanding assessments of the Big Five: Multitrait-multimethod analyses of convergent and discriminant validity across measurement occasion and type of observer. *Journal of Personality*, *72*, 845-876.

Bishop, Y. M. M. (1971). Effects of collapsing multidimensional contingency tables. *Biometrics*, *27*, 545-562.

Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Cambridge: MIT Press.

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

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*, 203-219.

Borsboom, D., Mellenbergh, G. J., & Van Heerden (2004). The concept of validity. *Psychological Review*, *111*, 1061-1071.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52, 345-370.

Browne, M. W. (1984). The decomposition of multitrait-multimethod matrices. *British Journal of Mathematical and Statistical Psychology, 37,* 1-21.

Burns, G. L., & Haynes, S. N. (2006). Clinical Psychology: Construct validation with multiple sources of information and multiple settings. In M. Eid, & E. Diener (Eds.), *Handbook of Multimethod Measurement in Psychology* (pp. 401-418). Washington, DC: American Psychological Association.

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

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology,* 43, 551-558.

Clogg, C. C. (1981). New developments in latent structure analysis. In D. J. Jackson, & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research* (pp. 215-246). Beverly Hills, C.A.: Sage.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Colvin, C. R. (1993a). Childhood antecedents of young adult judgeability. *Journal of Personality*, *61*, 611-635.

Colvin, C. R. (1993b). Judgeable people: Personality, behavior, and competing explanations. *Journal of Personality and Social Psychology, 64, 861-873.*

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *2,* 322-328.

Costello, A. J. (1973). The reliability of direct observations. *Bulletin of the British Psychological Society, 26,* 105-108.

Courvoisier, D. S., Nussbeck, F. W., Eid, M., Geiser, C., & Cole, D. A. (in press). Analyzing the convergent validity of states and traits: Development and application of multimethod latent state-trait models. *Psychological Assessment.*

Darroch, J. N., & McCloud, P. I. (1986). Category distinguishability and observer agreement. *Australian Journal of Statistics*, *28,* 371-388.

de Menezes, L. M. (1999). On fitting latent class models for binary data: The estimation of standard errors. *British Journal of Mathematical and Statistical Psychology*, *52*, 149-168.

Dillon, W. R., & Mulani, N. (1984). A probabilistic latent class model for assessing inter-judge reliability. *Multivariate Behavioral Research, 19*, 438-458.

Dominicus, A., Skrondal, A., Gjessing, H. K., Pederson, N. L., & Palmgren, J. (2006). Likelihood ratio test in behavioral genetics: problems and solutions. *Behavior Genetics, 36,* 331-340.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika, 65,* 241-261.

Eid, M., Langeheine, R., & Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis: A primer. *Journal of Cross-Cultural Psychology, 34,* 195-210.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp.283-299). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Geiser, C. (2004). *Die Multitrait-Multimethod-Analyse: Entwicklung neuer Modelle und ihre Anwendung in der Differentiellen und Diagnostischen Psychologie* [Multitrait-multimethod analysis: Development of new models and their applications in personality psychology and psychological assessment.]. University of Geneva: Report for the German Research Foundation.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CTC(M-1) model. *Psychological Methods*, 8, 38-60.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology,* 43, 543-549.

Fienberg, S. E. (1980). *The analysis of cross-classified categorical data*. Cambridge: MIT Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement between many raters. *Psychological Bulletin, 76,* 378-382.

Flink, C., & Park, B. (1991). Increasing consensus in trait judgments through outcome dependency. *Journal of Experimental Social Psychology, 27,* 453-467.

Forman, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association, 87,* 476 - 486.

Funder, D. C. (1995). On the accuracy of personality judgment: A realistic approach. *Psychological Review, 102,* 652-670.

Funder, D. C., & Dobroth, K. M. (1987). Differences between traits: Properties associated with interjudge agreement. *Journal of Personality and Social Psychology*, *52*, 409-418.

Funder, D. C., & West, S. G. (Eds.) (1993). Viewpoints on Personality: Consensus, self-other agreement and accuracy in judgments of personality. *Journal of Personality, 61(4)*. [Special Issue]

Galindo-Garre, F. J. K., & Vermunt, J. K. (2004). The order-restricted association model: Two estimation algorithms and issues in testing. *Psychometrika*, 69, 641-654.

Galindo-Garre, F. J. K., & Vermunt, J.K, (2005). Testing log-linear models with inequality constraints: A comparison of asymptotic, bootstrap, and posterior predictive p values. *Statistica Neerlandica*, 59, 82-94.

Galindo-Garre, F. J. K., & Vermunt, J.K, (2006). Avoiding boundary estimates in latent class analysis by Bayesian posterior mode estimation. *Behaviormetrika*, *33*, 43-59.

Gelfland, D. M., & Hartmann, D. P. (1975). *Child behavior analysis and therapy*. New York: Pergamon.

Goodman, L. A. (1973). The analysis of multidimensional contingency tables when some variables are posterior to others: A modified path analysis approach. *Biometrika*, *60*, 179-192.

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable: A modified latent structure approach. *American Journal of Sociology*, *79*, 1179-1259.

Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231.

Goodman, L. A. (1978). *Analyzing qualitative/categorical data. Loglinear models and latent structure analysis*. London: Addison-Wesley Publishing Company.

Haberman, S. J. (1976). Iterative scaling procedures for log-linear models or frequency data derived by indirect observation. *Proceedings of the American Statistical Association 1975*: *Statistical Computing Section*, 45-50.

Haberman, S. J. (1977). Product models for frequency tables involving indirect observation. *Annals of Statistics*, *5*, 1124-1147.

Haberman, S. J. (1978). *Analysis of qualitative data. Vol. 1. Introductory topics*. New York: Academic Press.

Haberman, S. J. (1979). *Analysis of qualitative data. Vol 2: New developments*. New York: Academic Press.

Haberman, S. J. (1988). A stabilized Newton-Raphson Algorithm for loglinear models for frequency tables derived by indirect observation. In C. C. Clogg (Ed.), *Sociological Methodology* (Vol. 18) (pp. 193-211). Washington, DC: American Sociological Association.

Hagenaars, J. A. (1990). *Categorical longitudinal data: Loglinear panel, trend, and cohort analysis*. Newbury Park, CA: Sage.

Hagenaars, J. A. (1993). *Loglinear models with latent variables*. Newbury Park, CA: Sage University Papers Series: Quantitative Applications in the Social Sciences, 07-094.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, 10 (1),* 103-116.

Hawkins, R. P., & Dotson, V. A. (1975). Reliability scores that delude: An Alice in Wonderland trip through the misleading characteristics of interobserver agreement scores in interval recording. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application* (pp. 359-376). Englewood Cliffs, NJ: Prentice-Hall.

Heinen, A. G. (1993). *Discrete latent variable models*. Tilburg: Tilburg University Press.

Hopkins, B. L., & Hermann, J. A. (1977). Evaluating interobserver reliability of interval data. *Journal of Applied Behavior Analysis, 10 (1),* 121-126.

House, A. E., House, B .J., & Campbell, M. B. (1981). Measures of interobserver agreement: Calculation formulas and distribution effects. *Journal of Behavioral Assessment*, *3*, 37-57.

Jansen, P. G. W., & Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51*, 69-91.

Johnson, L. C., & Bolstad, O. D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research. In L. A. Hamerlynck, L. C. Hardy, & E. J. Mash (Eds.), *Behavior change: Methodology, concepts, and practice* (pp. 7-67). Champaign, IL: Research Press.

Jöreskog, K. G. (1969). A general approach to confirmatory factor analysis. *Psychometrika, 34*, 183-202.

Jöreskog, K. G. (1973). A general method for estimating a linear structural equation system. In A. S. Goldeberger, & O. D. Duncen (Eds.), *Structural equation models in the social sciences* (pp. 83-112). New York: Seminar Press.

Kelly, M. B. (1977). A review of the observational data collection and reliability procedures reported in the Journal of Applied Behavior Analysis. *Journal of Applied Behavior Analysis, 10 (1),* 97-101.

Kenny, D. A. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12,* 247-252.

Kenny, D. A. (1979). *Correlation and causality*. New York: Wiley.

Kenny, D. A. (1995). The multitrait-multimethod matrix: Design, analysis, and conceptual issues. In S.T. Fiske, & P.E. Shrout (Eds.), *Personality research, methods, and theory: A festschrift honoring Donald W. Fiske* (pp.111-124). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kenny, D. A., & Kashy, D. A. (1992). The analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin*, *112*, 165-172.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* (159-174).

Langeheine, R. (1988). New developments in latent class theory. In R. Langeheine, & J. Rost (Eds.). *Latent trait and latent class models* (pp.77-108). New York: Plenum.

Langeheine, R., & Rost, J. (Eds.). (1988). *Latent trait and latent class models*. New York: Plenum.

Lazarsfeld, P. F. (1950a). The interpretation and mathematical foundation of latent structure analysis. In S. Stouffer (Ed.), *Measurement and Prediction* (pp. 362-472). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F. (1950b). The logical and mathematical foundation of latent structure analysis. In S. Stouffer (Ed.), *Measurement and Prediction* (pp. 362-472). Princeton, NJ: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mifflin.

Liebetrau, A. M. (1983). *Measures of association*. Beverly Hills: Sage Publications.

Lischetzke, T., & Eid, M. (2003). Is attention to feelings beneficial or detrimental to affective well-being? Mood regulation as a moderator variable. *Emotion, 3*, 361-377.

Magidson, J., & Vermunt, J. K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays, *Sociological Methodology, 31*, 223-264.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling. Concepts, issues, and applications* (pp. 177-198). Thousands Oaks, CA: Sage.

McCutcheon, A. C. (1987). *Latent class analysis*. Beverly Hills: Sage Publications.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741-749.

Mitchell, S. K. (1979). Interobserver agreement, reliability, and generalizability of data collected in observational studies. *Psychological Bulletin*, *86,* 376-390.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus User's guide.* Los Angeles: Muthén & Muthén.

Neyer, F. J. (2006). Informant assessment. In: M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 41-59), Washington, DC: American Psychological Association.

Nussbeck, F. W. (2002). Das Multitrait-Multimethod-True-Score-Regression-Modell - Eine Untersuchung zur Anwendbarkeit des Modells für kontinuierliche und ordinale Variablen in der differentialpsychologischen Emotionsforschung . Unveröffentlichte Diplomarbeit an der Universität Trier - FB I - Psychologie. [The Multitrait-Multimethod True-Score-Regression Model - An investigation of its applicability to models with response variables with metrical and ordered outcomes in emotional research, Thesis (Diploma) at the University of Trier, Germany].

Nussbeck, F.W. (2006). Assessing multimethod association with categorical variables. In: M. Eid, & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 223-247), Washington, DC: American Psychological Association.

Nussbeck, F. W., Eid, M., & Lischetzke, T. (2006). Analyzing MTMM data with SEM for ordinal variables applying the WLSMV-estimator: What is the sample size needed for valid results? *British Journal of Mathematical and Statistical Psychology, 59*, 195-213.

Nussbeck, F.W., Eid, M., Geiser, C., Courvoisier, D., & Lischetzke, T. (2008). A CTC(M-1) model for different types of raters. Manuscript submitted for publication.

Oort, F. J. (2008). Three-mode models for multitrait-multimethod data. Manuscript submitted for publication.

Ostendorf, F. (1990). *Sprache und Persönlichkeitsstruktur. Zur Validität des Fünf-Faktoren-Modells der Persönlichkeit* [Language and personality structure. On the validity of the five-factor model of personality]. Regensburg, Germany: Roderer.

Poznanski, E. O., & Mokros, H. B. (2005). *Children's depression rating scale revised* (*CDRS-R*). Los Angeles: Western Psychological Services.

Repp, A. C., Deitz, D. E. D., Boles, S. M., Deitz, S. M., & Repp, C. F. (1976). Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis, 9 (1),* 109-113.

Reynolds, H. T. (1977a). *Analysis of nominal data*. Beverly Hills: Sage Publications.

Reynolds, H. T. (1977b). *The analysis of cross-classifications*. New York: Free Press.

Roskam, E. E. (1995). Graded responses and joining categories: A rejoinder to Andrich` "Models for Measurement, precision, and nondichotomization of graded responses". *Psychometrika, 60,* 27-35.

Roskam, E. E., & Jansen, P. G. W. (1989). Conditions for Rasch-dichotomizability of the unidimensional polytomous Rasch model. *Psychometrika*, *54,* 317-333.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika, Monograph Supplement No. 17*.

Saris, W. E., & van Meurs, A. (1991). *Evaluation of measurement instruments by meta-analysis of multitrait-multimethod studies.* Amsterdam: North Holland.

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology*, *55*, 289-303.

Schuster, C., & Smith, D. A. (2002). Indexing systematic rater agreement with a latent class model. *Psychological Methods*, *7,* 384-395.

Schuster, C., & Smith, D. A. (2006). Estimating with a Latent Class model the reliability of nominal judgments upon which two raters agree. *Educational and Psychological Measurement*, *66*, 739-747.

Schwartz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *5*, 461-464.

Shadish, W.R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Co.

Sternberg, R. J. (1992). Psychological Bulletin`s top 10 “hit parade”. *Psychological Bulletin, 112,* 387-388.

Steyer, R., & Eid, M. (2001). *Messen und Testen* [*Measurement and Testing*]. Berlin: Springer.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Suen, H. K., Ary, D., & Ary, R. (1986). A note on the relationship among eight indices of interobserver agreement. *Behavioral Assessment, 8*, 301-303.

Thompson, W. D., & Walter, S. D. (1988a). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology, 41*, 949-958.

Thompson, W. D., & Walter, S. D. (1988b). Kappa and the concept of independent errors. *Journal of Clinical Epidemiology, 41*, 969-970.

Uebersax, J. S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101*, 140-146.

Uebersax, J. S., & Grove, W. M. (1990). Latent class analysis of diagnostic agreement. *Statistics in Medicine, 9*, 559-572.

van de Pol, F., Langeheine, R., & de Jong, W. (1996). *PANMARK 3. Users manual. Panel analysis using Markov chains: A latent class analysis program* [computer software manual]. Voorburg, the Netherlands: Statistics Netherlands.

Vermunt, J. K. (1996). Causal log-linear modelling with latent variables and missing data. In U. Engel, & J. Reinecke (Eds.*), Analysis of change: advanced techniques in panel data analysis*, (pp. 35-60). Berlin: Walter de Gruyter.

Vermunt, J. K. (1997a). *LEM: A general program for the analysis of categorical data*. Retrieved from http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html (downloaded 08/09/2006).

Vermunt, J. K. (1997b). *Log-linear models for event histories.* Advanced quantitative techniques in the social sciences series. Thousand Oakes: Sage Publications.

Vermunt J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33*, 213-39.

Vermunt J. K. (2005). Mixed-effects logistic regression models for indirectly observed outcome variables. *Multivariate Behavioral Research, 40*, 281-301.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research, 17,* 33-51.

Vermunt, J. K., & Magidson, J. (2000). *Latent GOLD 2.0 user's guide*. Belmont, MA: Statistical Innovations Inc.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced*. Belmont, MA: Statistical Innovations Inc.

Wickens, T. D., (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, *9*, 1-26.

Winship, C., & Mare, R. D. (1989). Log-linear models for missing data: A latent class approach. *Sociological Methodology*, *19,* 331-368.

Wothke, W., & Browne, W. W. (1990). The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrika, 55,* 255-262.

Zwick, R. (1988). Another look at interrater agreement. *Psychological Bulletin, 103,* 374-378.

List of Figures

Figure	Title	Page
2.1	The CT-model for three constructs with method-specific trait variables	16
4.1	Basic log-linear model with one latent variable (*NEUS*) for neuroticism	45
4.2	The loglinear-model with two latent variables for neuroticism and conscientiousness	67
5.1	Log-linear model with two latent variables	80
5.2	The relationship between the five latent rater agreement models	106
5.3	Log-linear model with two latent variables representing the latent construct Neuroticism	110
5.4	The relationship between the four latent rater agreement models for interchangeable raters	123
6.1	Categorical Multitrait-Multirater model for two traits measured by two raters	128

List of Tables

Table	Title	Page
1.1.1	*Artificial frequency table of two categorical variables*	8
2.1.1	*Multitrait-Multimethod matrix for three traits (T_1, T_2, and T_3) measured by three methods (M_1, M_2, and M_3)*	13
2.3.1	*Two cross-classifications of two ratings (artificial data)*	19
2.3.2	*Cross-classification of two variables*	26
2.3.4	*Effect coded parameters in the quasi-independence I model*	31
2.3.5	*Independence sub-table in the quasi-independence I model (ambiguous cases in Schuster & Smith, 2006)*	32
2.3.6	*Agreement (Reliability) sub-table in the quasi-independence I model (obvious cases in Schuster & Smith, 2006)*	33
4.1.1	*Three observed response patterns for the self-report data measuring neuroticism (extracted from the complete table in Appendix A)*	44
4.1.2	*Sample description with respect to the most frequently studied subjects*	56
4.1.3	*Sample description with respect to age*	57
4.1.4	*Time the target person S knows peers A and B*	57
4.1.5	*Frequency distribution of the analyzed Big-Five Items (self-report)*	59
4.1.6	*Goodness-of-fit coefficients of the log-linear model with 1, 2, and 3 latent categories*	61
4.1.7	*Multiplicative log-linear parameters of the log-linear model with three latent categories representing neuroticism (self-report)*	62
4.1.8	*Effect-parameters (category against the two others) for the two indicators "moody" and "self-doubtful" in the log-linear model with three latent categories representing neuroticism*	64

4.1.9	*Conditional response probabilities in the log-linear model with three latent categories representing neuroticism*	65
4.2.1	*Goodness-of-fit coefficients of the CT and independence CT model with two three- categorical latent variables*	72
4.2.2	*Log-linear parameters of the measurement model of the CT model; neuroticism*	73
4.2.3	*Conditional response probabilities of the manifest response categories for the construct neuroticism in the CT model*	74
4.2.4	*Log-linear parameters of the measurement model of the CT model; conscientiousness*	75
4.2.5	*Conditional response probabilities of the manifest response categories for the construct conscientiousness in the CT model*	77
4.2.6	*Cross classification of the estimated proportions of the two latent variables (neuroticism and conscientiousness) in the latent saturated CT model*	78
5.1.1	*Goodness-of-fit coefficients of the rater agreement models with three-categorical variables for structurally different raters*	91
5.1.2	*Conditional response probabilities of the manifest response categories for the construct neuroticism (NEUS) in the saturated latent rater agreement model for structurally different raters (self-report)*	93
5.1.3	*Conditional response probabilities of the manifest response categories for the construct neuroticism (NEUA) in the saturated latent rater agreement model for structurally different raters (peer report A)*	94
5.1.4	*Cross classification of the two latent variables (NEUS and NEUA) in the saturated latent rater agreement model for structurally different raters*	95
5.1.5	*Distinguishability indices for the saturated latent rater agreement model for structurally different raters*	96
5.1.6	*Cross classification of the two latent variables (NEUS and NEUA) in the independence model for structurally different raters*	97

5.1.7	*Cross classification of the two latent variables (NEUS and NEUA) in the quasi-independence I latent rater agreement model for structurally different raters*	99
5.1.8	*Cross classification of the two latent variables (NEUS and NEUA) in the quasi-independence II latent rater agreement model for structurally different raters*	100
5.1.9	*Cross classification of the two latent variables (NEUS and NEUA) in the quasi-symmetry latent rater agreement model for structurally different raters*	102
5.1.10	*Distinguishability indices for the two latent variables (NEUS and NEUA) in the quasi-symmetry latent rater agreement model for structurally different raters*	103
5.1.11	*Conditional response probabilities in the one-variable model for structurally different raters*	104
5.2.1	*Goodness-of-Fit coefficients of the different latent rater agreement models for interchangeable raters*	118
5.2.2	*Log-linear parameters of the measurement model of the latent quasi-independence-II latent rater agreement model*	119
5.2.3	*Conditional response probabilities of the manifest response categories in the latent quasi-independence-II latent rater agreement model*	121
5.2.4	*Latent joint distribution of the quasi-independence II latent rater agreement model*	122
6.1.1	*Extracted part of the latent joint distribution in the saturated CT MTMR model for three categorical latent variables with four-variable effects*	132
6.1.2	*Extracted part of the latent joint distribution in the saturated CT MTMR model for four three-categorical latent variables with three-variable interactions as highest order effects*	140
6.1.3	*Extracted part of the latent joint distribution in the saturated CT MTMR model with two-variable effects as highest order effects for different combinations of two categorical latent variables*	145

6.2.1	*Extracted part of the latent joint distribution in the saturated (symmetry) CT MTMR model for interchangeable raters*	154
6.2.2	*Extracted part of the latent joint distribution in the saturated CT MTMR model with three-variable effects as highest order effects for four three-categorical latent variables*	162
6.2.3	*Extracted part of the latent joint distribution in the saturated CT MTMR model with two-variable effects as highest order effects for different combinations of two categorical latent variables*	166
6.3.1	*Goodness-of-fit coefficients of the CT MTMR models for structurally different raters*	170
6.3.2	*Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with four-variable effects for structurally different raters*	172
6.3.3	*Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with three-variable effects as highest order interactions for structurally different raters*	176
6.3.4	*Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with two-variable effects as highest order interactions for structurally different raters*	179
6.3.5	*Cross-classification of expected proportions for the latent variables representing neuroticism*	181
6.3.6	*Distinguishability index and category-specific agreement rates for neuroticism*	181
6.3.7	*Cross-classification of expected proportions for the latent variables representing conscientiousness*	183
6.3.8	*Distinguishability index and category-specific agreement rates for conscientiousness*	184
6.3.9	*Cross-classification of expected proportions for the latent variables originating in the self-report*	185
6.3.10	*Cross-classification of expected proportions for the latent variables originating in the peer report*	185

6.3.11	*Cross-classification of expected proportions for neuroticism originating in the self-report and conscientiousness originating in the peer report*	186
6.3.12	*Cross-classification of expected proportions for neuroticism originating in the peer and conscientiousness originating in the self-report*	187
6.3.13	*Method bias type II for the self-report*	187
6.3.14	*Method bias type II for the peer report*	188
6.3.15	*Conditional response probabilities for neuroticism in the CT MTMR model with two-variable effects as highest order interactions for interchangeable raters*	194
6.3.16	*Conditional response probabilities for conscientiousness in the CT MTMR model with two-variable effects as highest order interactions for interchangeable raters*	195
6.3.17	*Cross-classification of the latent categories for neuroticism and conscientiousness in the CT MTMR Model with two-variable effects as highest order interactions for interchangeable raters*	196
6.3.18	*Cross-classification of expected proportions for the latent variables representing neuroticism*	198
6.3.19	*Distinguishability index and category-specific agreement rates for neuroticism*	198
6.3.20	*Cross-classification of expected proportions for the latent variables representing conscientiousness*	199
6.3.21	*Distinguishability index and category-specific agreement rates for conscientiousness*	199
6.3.22	*Cross-classification of expected proportions for the latent variables originating in one peer rater*	200
6.3.23	*Cross-classification of expected proportions for the latent variables of different constructs originating in different peer reports*	201
6.3.24	*Method bias type II for peer reports*	202

Appendix A: Data Description

The German Version of the Big-Five scale (Ostendorf, 1990)
Im folgenden finden Sie eine Reihe von Eigenschaftsbegriffen.
Kreuzen Sie bitte die Antwort an, die am ehesten auf Sie als Person zutrifft.

Ich bin

	überhaupt nicht				sehr
	1	2	3	4	5
kontaktfreudig	○	○	○	○	○
warmherzig	○	○	○	○	○
arbeitsam	○	○	○	○	○
verletzbar	○	○	○	○	○
klug	○	○	○	○	○
gesellig	○	○	○	○	○
fleißig	○	○	○	○	○
rücksichtsvoll	○	○	○	○	○
intelligent	○	○	○	○	○
empfindlich	○	○	○	○	○
	überhaupt nicht				sehr

Ich bin

	überhaupt nicht				sehr
	1	2	3	4	5
pflichtbewußt	○	○	○	○	○
launenhaft	○	○	○	○	○
lebhaft	○	○	○	○	○
kenntnisreich	○	○	○	○	○
gutmütig	○	○	○	○	○
temperamentvoll	○	○	○	○	○
hilfsbereit	○	○	○	○	○
geistreich	○	○	○	○	○
selbstzweiflerisch	○	○	○	○	○
strebsam	○	○	○	○	○
	überhaupt nicht				sehr

The peer report form exactly corresponds to the self-report form except for the pronomina used to describe the acting person. Change from German "ich" to "er / sie" and flexation of the verb, from 1st person singular to 3rd person singular.

A.1 Frequency Distributions of the Big-Five Items

Table A.1.1

Frequency distribution of the Big-Five Items (Ostendorf, 1990) of the self-report data

		categories			
German item	English item	little	middle	highly	total
arbeitssam	industrious	93	165	220	478
verletzbar	vulnerable	43	75	360	478
fleißig	diligent	116	159	203	478
empfindlich	sensitive	63	77	338	478
pflichtbewußt	dutiful	29	93	356	478
launenhaft	moody	179	130	169	478
selbstzweiflerisch	self-doubtful	121	88	269	478
strebsam	ambitious	122	150	206	478

Notes. Categories 1 and 2 as well as 3 and 4 of the original scale have been collapsed.

Table A.1.2

Frequency distribution of the Big-Five Items (Ostendorf, 1990) of peer report A

			categories		
German item	English item	little	middle	highly	total
kontaktfreudig	sociable	34	69	375	478
arbeitssam	industrious	78	127	273	478
verletzbar	vulnerable	76	141	261	478
gesellig	companionable	22	67	389	478
fleißig	diligent	78	134	266	478
empfindlich	sensitive	113	150	215	478
pflichtbewußt	dutiful	52	107	319	478
launenhaft	moody	258	126	94	478
lebhaft	vivacious	43	123	312	478
temperamentvoll	spirited	107	152	219	478
selbstzweiflerisch	self-doubtful	217	126	135	478
strebsam	ambitious	110	140	228	478

Notes. Categories 1 and 2 as well as 3 and 4 of the original scale have been collapsed.

Table A.1.3

Frequency distribution of the Big-Five Items (Ostendorf, 1990) of peer report B

			categories		
German item	English item	little	middle	highly	total
kontaktfreudig	sociable	37	64	377	478
arbeitssam	industrious	68	127	283	478
verletzbar	vulnerable	79	137	262	478
gesellig	companionable	19	67	392	478
fleißig	diligent	82	131	265	478
empfindlich	sensitive	105	154	219	478
pflichtbewußt	dutiful	50	97	331	478
launenhaft	moody	282	108	88	478
lebhaft	vivacious	47	106	325	478
temperamentvoll	spirited	109	126	243	478
selbstzweiflerisch	self-doubtful	218	131	129	478
strebsam	ambitious	101	141	236	478

Notes. Categories 1 and 2 as well as 3 and 4 of the original scale have been collapsed.

A.2 Response Patterns of the Self-Report Data for Neuroticism

Table A.2.1

Response patterns of the observed self-report data and their frequencies for the basic one-variable Latent-Class model

A vulnerable	*B* sensitive	*C* moody	*D* self-doubtful	frequency	relative frequency
1	1	1	1	8	.02
1	1	1	2	3	.01
1	1	1	3	2	.00
1	1	2	1	3	.01
1	1	2	2	0	
1	1	2	3	2	.00
1	1	3	1	1	.00
1	1	3	2	0	
1	1	3	3	1	.00
1	2	1	1	5	.01
1	2	1	2	2	.00
1	2	1	3	2	.00
1	2	2	1	1	.00
1	2	2	2	2	.00
1	2	2	3	0	
1	2	3	1	1	.00
1	2	3	2	1	.00
1	2	3	3	0	
1	3	1	1	1	.00
1	3	1	2	0	
1	3	1	3	4	.01
1	3	2	1	1	.00
1	3	2	2	1	.00
1	3	2	3	0	
1	3	3	1	1	.00
1	3	3	2	0	
1	3	3	3	1	.00
2	1	1	1	7	.01
2	1	1	2	4	.01
2	1	1	3	1	.00
2	1	2	1	2	.00
2	1	2	2	0	
2	1	2	3	1	.00
2	1	3	1	3	.01
2	1	3	2	0	
2	1	3	3	2	.00
2	2	1	1	10	.02
2	2	1	2	5	.01
2	2	1	3	5	.01
2	2	2	1	1	.00

Table continues…

Table continued

2	2	2	2	6	.01
2	2	2	3	1	.00
2	2	3	1	1	.00
2	2	3	2	1	.00
2	2	3	3	0	
2	3	1	1	2	.00
2	3	1	2	0	
2	3	1	3	4	.01
2	3	2	1	2	.00
2	3	2	2	3	.01
2	3	2	3	7	.01
2	3	3	1	3	.01
2	3	3	2	0	
2	3	3	3	4	.01
3	1	1	1	8	.02
3	1	1	2	5	.01
3	1	1	3	3	.01
3	1	2	1	1	.00
3	1	2	2	1	.00
3	1	2	3	3	.01
3	1	3	1	2	.00
3	1	3	2	0	
3	1	3	3	0	
3	2	1	1	4	.01
3	2	1	2	6	.01
3	2	1	3	8	.02
3	2	2	1	1	.00
3	2	2	2	4	.01
3	2	2	3	5	.01
3	2	3	1	1	.00
3	2	3	2	1	.00
3	2	3	3	3	.01
3	3	1	1	21	.04
3	3	1	2	9	.02
3	3	1	3	50	.10
3	3	2	1	18	.04
3	3	2	2	15	.03
3	3	2	3	49	.10
3	3	3	1	12	.03
3	3	3	2	19	.04
3	3	3	3	111	.23
			total	478	1

Notes. 1: non-neurotic response category; 2: neutral response category; 3: neurotic response category. Empty cells in the column representing the relative frequency indicate response patterns that have not been observed.

Appendix B: Collapsibility Theorem

Bishop (1971, p. 545) defined the “conditions… under which collapsing multidimensional contingency tables, by adding over variables, will affect the apparent interaction between the remaining variables”. Collapsing by adding over variables is also known as collapsing frequencies or categories of a variable or as collapsing arrays. A variable is considered collapsible if no interactions between variables, which remain in the reduced matrix, are changed compared to the effects in the full data matrix (see also .Bishop, Fienberg & Holland, 1975).

Theorem:

In a rectangular more-dimensional table a variable is collapsible with respect to the interaction between the other variables in a hierarchical model if and only if it is at least conditionally independent of all but one of the other variables given the last variable.

Proof:

See Bishop (1971) or Bishop. Fienberg, and Holland (1975).

Example:

Without loss of generality the simplest case of three variables is considered. The full additionally parameterized model reads as follows:

$$\ln(e_{ijk}) = \eta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}\,. \tag{B.0.1}$$

Without loss of generality, a model with one possible interaction absent is assumed. The interaction λ_{jk}^{BC} is chosen to be absent (which automatically leads to the absence of λ_{ijk}^{ABC}):

$$\ln(e_{ijk}) = \eta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC}\,. \tag{B.0.2}$$

The logarithms of the marginal sums may be written as:

$$\begin{aligned}
\ln(e_{ij+}) &= \eta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \\
\ln(e_{i+k}) &= \eta + \lambda_i^A + \lambda_k^C + \lambda_{ik}^{AC} \\
\ln(e_{ijk}) &= \eta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC}
\end{aligned} \tag{B.0.3}$$

The table is collapsible for the interaction of AC and AB.

Appendix C: Log-Linear Parameters of the Latent Rater Agreement Models for Structurally Different Raters

C.1 Saturated Latent Rater Agreement Model

Table C.1.1

Parameters of the measurement model of neuroticism of the saturated latent rater agreement model for structurally different raters (self report)

		one variable effect	two variables effect		
variable	manifest categories		*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	0.57	1.66	0.40	1.52
	2	0.17	7.92	4.49	0.03
	3	10.50	0.08	0.56	23.38
B (sensitive)	1	$8.51\ 10^{-54}$	$3.17\ 10^{53}$	$2.30\ 10^{52}$	$1.38\ 10^{-106}$*
	2	$6.37\ 10^{21}$	$3.80\ 10^{-22}$	$1.22\ 10^{-22}$	$2.15\ 10^{43}$*
	3	$1.84\ 10^{31}$	$8.32\ 10^{-33}$	$3.56\ 10^{-31}$	$3.38\ 10^{62}$*
C (moody)	1	1.07	2.53	1.10	0.36
	2	0.75	1.03	1.63	0.59
	3	1.25	0.38	0.56	4.70
D (doubtful)	1	0.47	3.54	1.85	0.15
	2	0.92	1.09	0.61	1.51
	3	2.31	0.26	0.89	4.35

Notes. * boundary values; *ns*: category of *NEUS*.

Table C.1.2

Parameters of the measurement model of neuroticism of the saturated latent rater agreement model for structurally different raters (peer report A)

variable	manifest categories	one variable effect	two variables effect1		
			na = 1	*na* = 2	*na* = 3
I (vulnerable)	1	0.17	2.48 10^{3}	1.75	2.29 10^{-4}*
	2	0.43	603.47 10^{53}	3.99	4.15 10^{-4}*
	3	13.57	6.69 10^{-7}*	0.14	1.04 10^{7}*
J (sensitive)	1	3.68 10-^{5}	4.37 10^{10}	1.57 10^{4}	1.45 10^{-15}
	2	1.04 10^{5}	5.32	1.46 10^{-5}	1.28 10^{4}
	3	0.26	4.31 10^{-12}*	4.33	5.36 10^{10}
K (moody)	1	1.97	1.77	1.07	0.53
	2	0.85	0.83	1.35	0.90
	3	0.59	0.69	0.69	2.11
L (doubtful)	1	1.53	2.24	1.00	0.44
	2	0.81	0.78	1.24	1.04
	3	0.81	0.57	0.80	2.17

Notes. * boundary values; n*a*: category of *NEUA*.

C.2 Independence Latent Rater Agreement Model

Table C.2.1

Parameters of the measurement model of neuroticism of the independence latent rater agreement model for structurally different raters (self report)

		one variable effect	two variables effect		
variable	manifest categories		*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	2.74 10^{-37}	3.22 10^{36}	8.97 10^{35}	3.47 10^{-73}*
	2	4.20 10^{17}	3.00 10^{-18}	1.66 10^{-18}	2.01 10^{35}
	3	8.69 10^{18}	1.03 10^{-19}	6.73 10^{-19}	1.43 10^{37}
B (sensitive)	1	1.07 10^{6}	5.76 10^{6}	1.74 10^{-7}	0.00**
	2	5.76 103	1.06 10^{9}	1.14 10^{-4}	8.24 10^{-6}
	3	1.66 10^{-8}*	1.60 10^{-18}	4.92 10^{8}	1.27 10^{9}
C (moody)	1	0.73	3.65	1.60	0.17
	2	0.36	2.25	3.36	0.13
	3	3.78	0.12	0.19	44.03
D (doubtful)	1	0.47	3.37	1.83	0.16
	2	0.91	1.11	0.59	1.52
	3	2.36	0.27	0.93	4.04

Notes. * boundary values; **: zero fitted margin; *ns*: category of *NEUS*..

Table C.2.2

Parameters of the measurement model of neuroticism of the latent rater agreement model for structurally different raters (peer report A)

		one variable effect	two variables effect		
variable	manifest categories		*na* = 1	*na* = 2	*na* = 3
I (vulnerable)	1	2.11	16.71	0.00*	57.14
	2	0.10	202.26	198.50*	2.49 10^{-5*}
	3	4.58	2.96 10^{-4}	4.80*	702.64*
J (sensitive)	1	1.14 10^{-5}	3.30 10^{5}	5.34 10^{4}	5.68 10^{-11}
	2	324.27	0.01	0.00	3.70 10^{4}
	3	269.13	5.51 10^{-4}	0.00	4.75 10^{5}
K (moody)	1	1.95	1.53	1.24	0.53
	2	0.90	0.82	1.40	0.88
	3	0.57	0.80	0.58	2.18
L (doubtful)	1	1.55	2.09	0.89	0.54
	2	0.82	0.82	1.25	0.98
	3	0.79	0.59	0.90	1.90

Notes. * boundary values; *na*: category of *NEUA*.

Table C.2.3

Conditional probabilities of the manifest response categories for the construct neuroticism (NEUS) in the quasi-independence I latent rater agreement model for structurally different raters (self-report)

		latent status		
variable	manifest categories	*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	.31	.04	.01
	2	.42	.12	.00
	3	.27	.84	.99
B (sensitive)	1	.51	.03	.00*
	2	.47	.11	.00
	3	.01	.86	1.00
C (moody)	1	.69	.38	.11
	2	.20	.33	.13
	3	.12	.23	.76
D (doubtful)	1	.50	.28	.00
	2	.31	.15	.13
	3	.18	.57	.87

Notes. * boundary values; *ns*: categories of *NEUS*.

Table C.2.4

Conditional probabilities of the manifest response categories for the construct neuroticism (NEUA) in the quasi-independence I latent rater agreement model for structurally different raters (peer report A)

		latent status		
variable	manifest categories	*na* = 1	*na* = 2	*na* = 3
I (vulnerable)	1	.58	.09	.00*
	2	.39	.44	.00*
	3	.03	.47	1.00*
J (sensitive)	1	.78	.16	.00*
	2	.22	.50	.08
	3	.00*	.34	.92
K (moody)	1	.76	.57	.34
	2	.15	.32	.25
	3	.09	.11	.40
L (doubtful)	1	.75	.48	.22
	2	.15	.31	.26
	3	.11	.21	.52

Notes. * boundary values; *na*: categories of *NEUA*.

C.3 Quasi-Independence I Latent Rater Agreement Model

Table C.3.1

Parameters of the measurement model of neuroticism of the quasi-independence I latent rater agreement model for structurally different raters (self report)

variable	manifest categories	one variable effect	two variables effect		
			ns = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	0.45	2.10	0.53	0.90
	2	0.32	4.00	2.41	0.10
	3	6.88	0.12	0.79	10.68
B (sensitive)	1	5.74 $10^{-40\,*}$	6.06 $10^{39\,*}$	3.64 $10^{38\,*}$	4.54 $10^{-79\,*}$
	2	5.67 $10^{18\,*}$	5.62 $10^{-19\,*}$	1.37 $10^{-19\,*}$	1.30 $10^{37\,*}$
	3	3.07 $10^{20\,*}$	2.93 $10^{-22\,*}$	2.01 $10^{-20\,*}$	1.69 $10^{41\,*}$
C (moody)	1	1.17	2.33	1.00	0.43
	2	0.82	0.97	1.47	0.71
	3	1.05	0.44	0.68	3.30
D (doubtful)	1	0.04	42.17	24.94	9.51 10^{-4}
	2	3.26	0.31	0.16	19.67
	3	7.91	0.08	0.25	53.47

Notes. * boundary values; *ns*: category of *NEUS.*

Table C.3.2

Parameters of the measurement model of neuroticism of the quasi-independence I latent rater agreement model for structurally different raters (peer report A)

		one variable effect	two variables effect1		
variable	manifest categories		*na* = 1	*na* = 2	*na* = 3
	1	0.07	44.07	4.45*	6.33 10^{-10}
I (vulnerable)	2	0.11	19.30	14.85*	574.75
	3	119.83	0.00	0.64*	2.75 10^{6}
	1	8.00 10^{-4}	2.32 10^{6}	681.21	6.33 10^{-10}
J (sensitive)	2	711.33	0.75	0.00	574.75
	3	1.76	5.72 10^{-7}	0.64	2.75 10^{6}
	1	1.97	1.74	1.08	0.53
K (moody)	2	0.85	0.81	1.37	0.90
	3	0.60	0.71	0.68	2.07
	1	1.53	2.15	1.00	0.47
L (doubtful)	2	0.82	0.80	1.22	1.02
	3	0.80	0.58	0.82	2.10

Notes. * boundary values; n*a*: category of *NEUA*.

Table C.3.3

Conditional probabilities of the manifest response categories for the construct neuroticism (NEUS) in the quasi-independence I latent rater agreement model for structurally different raters (self-report)

		latent status		
variable	manifest categories	*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	.31	.04	.01
	2	.42	.12	.00
	3	.27	.84	.99
B (sensitive)	1	.51	.03	.00*
	2	.47	.11	.00
	3	.01	.86	1.00
C (moody)	1	.69	.38	.11
	2	.20	.33	.13
	3	.12	.23	.76
D (doubtful)	1	.50	.28	.00
	2	.31	.15	.13
	3	.18	.57	.87

Notes. * boundary values; *ns*: categories of *NEUS*.

Table C.3.4

Conditional probabilities of the manifest response categories for the construct neuroticism (NEUA) in the quasi-independence I latent rater agreement model for structurally different raters (peer report A)

		latent status		
variable	manifest categories	*na* = 1	*na* = 2	*na* = 3
I (vulnerable)	1	.58	.09	.00*
	2	.39	.44	.00*
	3	.03	.47	1.00*
J (sensitive)	1	.78	.16	.00*
	2	.22	.50	.08
	3	.00*	.34	.92
K (moody)	1	.76	.57	.34
	2	.15	.32	.25
	3	.09	.11	.40
L (doubtful)	1	.75	.48	.22
	2	.15	.31	.26
	3	.11	.21	.52

Notes. * boundary values; *na*: categories of *NEUA*.

C.4 Quasi-Independence II Latent Rater Agreement Model

Table C.4.1

Parameters of the measurement model of neuroticism of the quasi-independence II latent rater agreement model for structurally different raters (self report)

		one variable effect	two variables effect		
variable	manifest categories		*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	0.36	2.68	0.41	0.90
	2	0.66	1.96	1.53	0.33
	3	4.29	0.19	1.58	3.32
B (sensitive)	1	5.20 10^{-55}	4.03 10^{54}	3.99 10^{53}	6.23 10^{-10}9*
	2	9.35 10^{25}	1.99 10^{26}	9.54 10^{-27}	5.27 10^{51}
	3	2.06 10^{28}	1.25 10^{-29}	2.63 10^{-28}	3.04 10^{56}
C (moody)	1	1.27	2.06	0.98	0.50
	2	0.84	0.92	1.64	0.66
	3	0.93	0.53	0.62	3.05
D (doubtful)	1	0.75	2.20	1.29	0.35
	2	0.75	1.32	0.75	1.01
	3	1.79	0.34	1.04	2.81

Notes. * boundary values; *ns*: category of *NEUS.*

Table C.4.2

Parameters of the measurement model of neuroticism of the quasi-independence II latent rater agreement model for structurally different raters (peer report A)

variable	manifest categories	one variable effect	two variables effect		
			na = 1	*na* = 2	*na* = 3
I (vulnerable)	1	2.84	0.91	0.09	12.70
	2	0.04	47.59	56.14	$3.74\ 10^{-4}$
	3	9.92	0.02	0.21	210.36
J (sensitive)	1	$3.00\ 10^{-4}$	$1.53\ 10^{5}$	$1.80\ 10^{3}$	$3.63\ 10^{-9}$*
	2	201.06	0.08	0.01	$1.46\ 10^{3}$
	3	16.58	$7.83\ 10^{-5}$	0.07	$1.89\ 10^{5}$
K (moody)	1	1.97	1.65	1.13	0.53
	2	0.87	0.80	1.43	0.87
	3	0.58	0.75	0.62	2.15
L (doubtful)	1	1.54	2.07	0.95	0.51
	2	0.82	0.80	1.24	0.99
	3	0.80	0.60	0.84	1.98

Notes. * boundary values; n*a*: category of *NEUA*.

Table C.4.3

Conditional probabilities of the manifest response categories for the construct neuroticism NEUS) in the quasi-independence II latent rater agreement model for structurally different raters (self-report)

		latent status		
variable	manifest categories	*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	.31	.02	.02
	2	.42	.13	.01
	3	.27	.86	.96
B (sensitive)	1	.50	.03	.00*
	2	.41	.14	.00*
	3	.06	.83	1.00*
C (moody)	1	.67	.39	.16
	2	.20	.43	.14
	3	.13	.18	.70
D (doubtful)	1	.51	.29	.04
	2	.30	.17	.13
	3	.19	.55	.83

Notes. * boundary values; *ns*: categories of *NEUS*.

Table C.4.4

Conditional probabilities of the manifest response categories for the construct neuroticism (NEUA) in the quasi-independence II latent rater agreement model for structurally different raters (peer report A)

		latent status		
variable	manifest categories	*na* = 1	*na* = 2	*na* = 3
I (vulnerable)	1	.57	.06	.02
	2	.37	.47	.00*
	3	.05	.47	.98
J (sensitive)	1	.73	.16	.00*
	2	.27	.50	.09
	3	.00	.34	.91
K (moody)	1	.74	.58	.34
	2	.16	.32	.25
	3	.10	.09	.41
L (doubtful)	1	.74	.46	.25
	2	.15	.32	.26
	3	.11	.21	.50

Notes. * boundary values; *na*: categories of *NEUA*.

C.5 Quasi-Symmetry Latent Rater Agreement Model

Table C.5.1

Parameters of the measurement model of neuroticism of the quasi-symmetry latent rater agreement model for structurally different raters (self report)

		one variable effect	two variables effect		
variable	manifest categories		*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	0.45	2.09	0.56	0.91
	2	0.32	4.04	2.43*	0.10
	3	6.91	0.12	0.78	10.78
B (sensitive)	1	4.42 10^{-40}	7.87 10^{39}	4.72 10^{38}	2.69 10^{-79}*
	2	6.67 10^{18}	4.78 10^{-19}	1.16 10^{-19}	1.80 10^{37}
	3	3.39 10^{20}	2.66 10^{-22}	1.82 10^{-20}	2.06 10^{41}
C (moody)	1	1.17	2.33	1.00	0.43
	2	0.82	0.97	1.47	0.71
	3	1.05	0.44	0.69	3.30
D (doubtful)	1	0.03	53.01	31.36	6.02 10^{-4}
	2	3.66	0.28	0.15	27.73
	3	8.87	0.07	0.22	67.22

Notes. * boundary values; *ns*: category of *NEUS*.

Table C.5.2

Parameters of the measurement model of neuroticism of the quasi-symmetry latent rater agreement model for structurally different raters (peer report A)

variable	manifest categories	one variable effect	two variables effect		
			na = 1	*na* = 2	*na* = 3
I (vulnerable)	1	0.07	43.41	4.38	0.01*
	2	0.13	17.15	13.20	0.04
	3	104.94	0.00	0.02	$4.31\ 10^{4}$
J (sensitive)	1	0.00	$1.53\ 10^{6}$	521.43	$1.25\ 10^{-9}$*
	2	577.28	0.80	0.00	439.96
	3	1.66	$8.20\ 10^{-7}$	0.67	$1.81\ 10^{6}$
K (moody)	1	1.97	1.74	1.08	0.53
	2	0.85	0.81	1.37	0.90
	3	0.60	0.71	0.68	2.07
L (doubtful)	1	1.53	2.15	1.00	0.47
	2	0.82	0.80	1.22	1.02
	3	0.80	0.58	0.82	2.10

Notes. * boundary values; n*a*: category of *NEUA*.

Table C.5.3

Conditional probabilities of the manifest response categories for the construct neuroticism in the quasi-symmetry latent rater agreement model for structurally different raters (self-report)

		latent status		
variable	manifest categories	*ns* = 1	*ns* = 2	*ns* = 3
A (vulnerable)	1	.31	.04	.01
	2	.42	.12	.00
	3	.27	.84	.99
B (sensitive)	1	.51	.03	.00*
	2	.47	.11	.00
	3	.01	.86	1.00
C (moody)	1	.69	.38	.11
	2	.20	.39	.13
	3	.12	.23	.76
D (doubtful)	1	.50	.28	.00
	2	.31	.15	.13
	3	.18	.57	.87

Notes. * boundary values; *ns*: categories of *NEUS*.

Table C.5.4

Conditional probabilities of the manifest response categories for the construct neuroticism (NEUA) in the quasi-symmetry latent rater agreement model for structurally different raters (peer report A)

		latent status		
variable	manifest categories	*na* = 1	*na* = 2	*na* = 3
I (vulnerable)	1	.58	.09	.00*
	2	.39	.44	.00
	3	.03	.47	1.00
J (sensitive)	1	.77	.16	.00*
	2	.22	.50	.08
	3	.00	.34	.92
K (moody)	1	.76	.57	.34
	2	.15	.31	.25
	3	.09	.11	.40
L (doubtful)	1	.75	.48	.22
	2	.15	.31	.26
	3	.11	.21	.52

Notes. * boundary values; *na*: categories of *NEUA*.

Appendix E: Loglinear Parameters of the CT MTMR Model with Two-Variable Effects as Highest Order Interactions

The interpretation of the log-linear parameters should be carried out very cautiously because LEM encounters difficulties estimating large log-linear models with latent variables.

E.1: Loglinear Parameters of the Model for Structurally Different Raters

Table E.1.1

Cross-classification of the log-linear two-variable effects for the construct neuroticism

		NEUA		
	1	2	3	
ns = 1	2.08	0.89	0.54	0.73
ns = 2	0.59	1.43	1.18	1.28
ns = 3	0.81	0.79	1.58	1.08
	0.95	1.06	0.99	

Notes. ^{s} indicates parameters with *z*-values larger than 2.00. ^{b} indicates boundary values.

Table E.1.2

Cross-classification of the log-linear two-variable effects for the construct conscientiousness

		CONA		
	1	2	3	
cs = 1	1.87	1.14	0.47	0.89
cs = 2	1.10	1.09	0.83	1.26
cs = 3	0.49	0.81	2.55	0.89
	0.54	1.24	1.49	

Notes. ^{s} indicates parameters with *z*-values larger than 2.00. ^{b} indicates boundary values.

Table E.1.3

Cross-classification of the log-linear two-variable effects for the self-report

	CONS		
	1	2	3
ns = 1	0.96	1.04	0.99
ns = 2	0.84	1.12	1.06
ns = 3	1.23	0.85	0.95

Notes. ^{s} indicates parameter values with *z*-values larger than 2.00.

Table E.1.4

Cross-classification of the log-linear two-variable effects for the peer report

	CONA		
	1	2	3
na = 1	1.83^{s}	1.11	0.49^{s}
na = 2	0.62	0.89	1.83
na = 3	0.89	1.01	1.11

Notes. ^{s} indicates parameters with *z*-values larger than 2.00. ^{b} indicates boundary values.

Table E.1.5

Cross-classification of the log-linear two-variable effects for neuroticism in the self-report and conscientiousness in the peer report

	CONA		
	1	2	3
ns = 1	0.92	0.93	1.16
ns = 2	1.17	0.94	0.91
ns = 3	0.93	1.14	0.95

Table E.1.6

Cross-classification of the log-linear two-variable effects for neuroticism in the peer report and conscientiousness in the self-report

	CONS		
	1	2	3
na = 1	1.18	0.90	0.94
na = 2	1.01	0.99	1.00
na = 3	0.84	1.12	1.06

E.2: Loglinear Parameters of the Model for Interchangeable Raters

Table E.2.1

Cross-classification of the log-linear two-variable for neuroticism

	NEUB			
	1	2	3	
na = 1	7.24[s]	5.81[s]	1.22[b]	0.13[s]
na = 2	5.81[s]	7.96[s]	1.21[b]	0.12[s]
na = 3	1.22[b]	1.21[b]	1	1
	0.13[s]	0.12[s]	1	

Notes. [s] indicates parameters with *z*-values larger than 2.00. [b] indicates boundary values.

Table E.2.2

Cross-classification of the log-linear two-variable effects for conscientiousness

	CONB			
	1	2	3	
ca = 1	3.20[s]	3.03[s]	0.39[b]	0.40[s]
ca = 2	3.03[s]	4.03[s]	0.89[b]	0.57[s]
ca = 3	0.39[b]	0.89[b]	1	1
	0.40[s]	0.57[s]	1	

Notes. [s] indicates parameters with *z*-values larger than 2.00. [b] indicates boundary values.

Table E.2.3

Cross-classification of the log-linear two-variable effects across traits

	CONA		
	1	2	3
na = 1	2.61 ^{s}	1.75	2.40 ^{b}
na = 2	0.74	1.81	4.40 ^{b}
na = 3	0.53 ^{b}	0.51 ^{b}	1

Notes. Due to the equality restrictions the same parameter values are found for the associations of *NEUB* and *CONB*.

Table E.2.4

Cross-classification of the log-linear two-variable effects across raters

	CONB		
	1	2	3
na = 1	0.94	0.53 ^{s}	0.67
na = 2	0.71	0.69	0.77 ^{b}
na = 3	0.62 ^{b}	0.77 ^{b}	1

Notes. Due to the equality restrictions the same parameter values are found for the associations of *NEUB* and *CONA*.

Appendix F: Input Files

F.1: Input Files for Rater Agreement Models

<u>Saturated Model</u>

```
man 2
dim 3 3
lab A B
mod {AB}
```

<u>Independence Model</u>

```
man 2
dim 3 3
lab A B
mod {A,B}
```

<u>Quasi-independence I Model</u>

```
man 2
dim 3 3
lab A B
mod {spe (AB, 5a)}
```

<u>Quasi-Independence II Model</u>

```
man 2
dim 3 3
lab A B
mod {A,B fac(AB,1)}
des [1 0 0
     0 1 0
     0 0 1]
```

<u>Quasi-symmetry Model</u>

```
man 2
dim 3 3
lab A B
mod {A,B, fac(AB,3)}
des [1 2 0
     2 3 0
     0 0 0]
```

<u>Symmetry Model</u>

```
man 2
dim 3 3
lab A B
mod {spe (AB, 3a)}
```

F.2: Input File for the Log-Linear Models with One Latent Variable and the model with two latent variables

Log-linear Model with One Latent Variable

```
lat 1
man 4
dim 3 3 3 3 3
lab X A B C D
mod  {X,X.A,X.B,X.C,X.D}
```

Log-linear Model with Two Latent Variables

```
      lat 2
      man 8
      dim 3 3 3 3 3 3 3 3 3 3
      lab NE CO E A F B G C D H
      mod  NE.CO {NE,CO}

     A|NE {A.NE}
     B|NE {B.NE}
     C|NE {C.NE}
     D|NE {D.NE}
     E|CO {E.CO}
     F|CO {F.CO}
     G|CO {G.CO}
     H|CO {H.CO}
```

F.3: Input Files for the Latent Rater Agreement Models

Saturated Model for Structurally Different Raters

```
      lat 2
      man 8
      dim 3 3 3 3 3 3 3 3 3 3
      lab  SN AN A B C D I J K L
      mod  SN.AN

     A|SN {A.SN}
     B|SN {B.SN}
     C|SN {C.SN}
     D|SN {D.SN}
     I|AN {I.AN}
     J|AN {J.AN}
     K|AN {K.AN}
     L|AN {L.AN}

 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

Quasi-symmetry Model for Structurally Different Raters

```
      lat 2
      man 8
      dim 3 3 3 3 3 3 3 3 3 3
      lab  SN AN A B C D I J K L
      mod  SN.AN {SN,AN, fac(SN.AN,5)}
```

```
     A|SN {A.SN}
     B|SN {B.SN}
     C|SN {C.SN}
     D|SN {D.SN}
     I|AN {I.AN}
     J|AN {J.AN}
     K|AN {K.AN}
     L|AN {L.AN}

 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]

des [      1 2 3
           2 4 5
           3 5 0]
```

Quasi-independence I Model for Structurally Different Raters

```
      lat 2
      man 8
      dim 3 3 3 3 3 3 3 3 3 3
      lab  SN AN A B C D I J K L
      mod  SN.AN {SN,AN, spe(SN.AN,5a)}

     A|SN {A.SN}
     B|SN {B.SN}
     C|SN {C.SN}
     D|SN {D.SN}
     I|AN {I.AN}
     J|AN {J.AN}
     K|AN {K.AN}
     L|AN {L.AN}

 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

Quasi-independence II Model for Structurally Different Raters

```
      lat 2
      man 8
      dim 3 3 3 3 3 3 3 3 3 3
      lab  SN AN A B C D I J K L
      mod  SN.AN {SN,AN, fac(SN.AN,1)}

     A|SN {A.SN}
     B|SN {B.SN}
     C|SN {C.SN}
     D|SN {D.SN}
     I|AN {I.AN}
     J|AN {J.AN}
     K|AN {K.AN}
     L|AN {L.AN}

 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]

des  [     1 0 0
           0 1 0
           0 0 1]
```

Independence Model for Structurally Different Raters

```
      lat 2
      man 8
      dim 3 3 3 3 3 3 3 3 3 3
      lab  SN AN A B C D I J K L
      mod  SN,AN

     A|SN {A.SN}
     B|SN {B.SN}
     C|SN {C.SN}
     D|SN {D.SN}
     I|AN {I.AN}
     J|AN {J.AN}
     K|AN {K.AN}
     L|AN {L.AN}

 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

Symmetry Model for Interchangeable Raters

```
        lat 2
        man 8
        dim 3 3 3 3 3 3 3 3 3 3
        lab X Y E F G H I J K L
        mod  XY {fac(X,Y,2),fac(XY,8)}

     E|X {EX}
     F|X {FX}
     G|X {GX}
     H|X {HX}
     I|Y eq1 E|X
     J|Y eq1 F|X
     K|Y eq1 G|X
     L|Y eq1 H|X

des [      1 2 0 * to model the marginal of X
           1 2 0 * to model the marginal of y
           1 2 3
           4 5 6
           7 8 0]

sta E|X [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

Quasi-independence I Model for Interchangeable Raters

```
        lat 2
        man 8
        dim 3 3 3 3 3 3 3 3 3 3
        lab X Y E F G H I J K L
        mod  XY {fac(X,Y,2),fac(XY,3)}

     E|X {EX}
     F|X {FX}
     G|X {GX}
     H|X {HX}
     I|Y eq1 E|X
     J|Y eq1 F|X
     K|Y eq1 G|X
     L|Y eq1 H|X

des [      1 2 0 * to model the marginal of X
           1 2 0 * to model the marginal of y
           1 0 0
           0 2 0
           0 0 3]

sta E|X [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

Quasi-independence II Model for Interchangeable Raters

```
        lat 2
        man 8
        dim 3 3 3 3 3 3 3 3 3 3
        lab X Y E F G H I J K L
        mod  XY {fac(X,Y,2),fac(XY,1)}

     E|X {EX}
     F|X {FX}
     G|X {GX}
     H|X {HX}
     I|Y eq1 E|X
     J|Y eq1 F|X
     K|Y eq1 G|X
     L|Y eq1 H|X

des [      1 2 0 * to model the marginal of X
           1 2 0 * to model the marginal of y
           1 0 0
           0 1 0
           0 0 1]

sta E|X [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

Independence Model for Interchangeable Raters

```
        lat 2
        man 8
        dim 3 3 3 3 3 3 3 3 3 3
        lab X Y E F G H I J K L
        mod  XY {fac(X,Y,2)}

     E|X {EX}
     F|X {FX}
     G|X {GX}
     H|X {HX}
     I|Y eq1 E|X
     J|Y eq1 F|X
     K|Y eq1 G|X
     L|Y eq1 H|X

des [      1 2 0 * to model the marginal of X
           1 2 0] * to model the marginal of y

sta E|X [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

F.4: Input Files for the CT MTMR Models for Structurally Different and Interchangeable Raters

Saturated CT MTMR Model for Structurally Different Raters

```
      lat 4
      man 16
      dim 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
      lab  SN SC AN AC E A F B G C D H M I N J O K L P
      mod  SN.SC.AN.AC

     A|SN {SN.A}
     B|SN {SN.B}
     C|SN {SN.C}
     D|SN {SN.D}
     E|SC {E.SC}
     F|SC {F.SC}
     G|SC {G.SC}
     H|SC {H.SC}
     I|AN {I.AN}
     J|AN {J.AN}
     K|AN {K.AN}
     L|AN {L.AN}
     M|AC {M.AC}
     N|AC {N.AC}
     O|AC {O.AC}
     P|AC {P.AC}

 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta E|SC [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta M|AC [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

CT MTMR Model with Three-Variable Effects as Highest Order Interactions for Structurally Different Raters

```
      lat 4
      man 16
      dim 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
      lab  SN SC AN AC E A F B G C D H M I N J O K L P
      mod  SN.SC.AN.AC {SN.SC.AN,SN.SC.AC,SN.AN.AC,SC.AN.AC}

     A|SN {SN.A}
     B|SN {SN.B}
     C|SN {SN.C}
     D|SN {SN.D}
     E|SC {SC.E}
     F|SC {SC.F}
     G|SC {SC.G}
     H|SC {SC.H}
     I|AN {I.AN}
     J|AN {J.AN}
     K|AN {K.AN}
     L|AN {L.AN}
     M|AC {M.AC}
     N|AC {N.AC}
     O|AC {O.AC}
     P|AC {P.AC}
```

```
 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta E|SC [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta M|AC [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

CT MTMR Model with Two-Variable Effects as Highest Order Interactions for Structurally Different Raters

```
      lat 4
      man 16
      dim 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
      lab  SN SC AN AC E A F B G C D H M I N J O K L P
      mod  SN.SC.AN.AC {SN.SC,SN.AN,SN.AC,SC.AN,SC.AC,AN.AC}


     A|SN {SN.A}
     B|SN {SN.B}
     C|SN {SN.C}
     D|SN {SN.D}
     E|SC {SC.E}
     F|SC {SC.F}
     G|SC {SC.G}
     H|SC {SC.H}
     I|AN {AN.I}
     J|AN {AN.J}
     K|AN {AN.K}
     L|AN {AN.L}
     M|AC {AC.M}
     N|AC {AC.N}
     O|AC {AC.O}
     P|AC {AC.P}

 sta A|SN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta E|SC [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta I|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta M|AC [.90 .05 .05 .05 .90 .05 .05 .05 .90]
```

CT MTMR Model with Two-Variable Effects as Highest Order Interactions for Interchangeable Raters

```
      lat 4
      man 16
      dim 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
      lab  AN AC BN BC E F G H I J K L M N O P Q R S T
      mod  AN.AC.BN.BC {fac(AN,BN,2), fac(AN.BN,5), fac(AC,BC,2),
fac(AC.BC,5), fac(AN.AC,BN.BC,8), fac(AN.BC,AC.BN,8),
fac(AN.AC.BN,AN.BN.BC,0), fac(AC.BN.BC,AN.AC.BC,0),
fac(AN.AC.BN.BC,0)}

     F|AN {AN.F}
     H|AN {AN.H}
     J|AN {AN.J}
     K|AN {AN.K}
     E|AC {E.AC}
     G|AC {G.AC}
     I|AC {I.AC}
     L|AC {L.AC}
     N|BN eq1 F|AN
     P|BN eq1 H|AN
     R|BN eq1 J|AN
     S|BN eq1 K|AN
     M|BC eq1 E|AC
     O|BC eq1 G|AC
     Q|BC eq1 I|AC
     T|BC eq1 L|AC

 sta F|AN [.90 .05 .05 .05 .90 .05 .05 .05 .90]
 sta E|AC [.90 .05 .05 .05 .90 .05 .05 .05 .90]

des [1     2     0 *AN,BN
     1     2     0
     1     2     3 *AN.BN
     2     4     5
     3     5     0
     1     2     0 *AC,BC
     1     2     0
     1     2     3 *AC.BC
     2     4     5
     3     5     0
     1     2     3 *AN.AC = BN.BC
     4     5     6
     7     8     0
     1     2     3 *BN.BC = AN.AC
     4     5     6
     7     8     0
     1     2     3 *AN.BC = AC.BN
     4     5     6
     7     8     0
     1     4     7 *AC.BN = AN.BC
     2     5     8
     3     6     0

    0  0  0  0  0  0  0  0  0 *AN.AC.BN = AN.BN.BC
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0 *AN.BN.BC = AN.AC.BN
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0 * AN.AC.BC = AC.BN.BC
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0 *AC.BN.BC= AN.AC.BC
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0  *AN.AC.BN.BC
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0
    0  0  0  0  0  0  0  0  0]
```

Appendix G in German Language (Anhang in deutscher Sprache)

G.1: Zusammenfassung in deutscher Sprache

In der vorliegenden Arbeit werden Modelle zur Erfassung der Beurteilerübereinstimmung für latente kategoriale Variablen (latent rater agreement models) und Multitrait-Multirater Modelle definiert, um die konvergente und diskriminante Validität von kategorialen Daten messfehlerbereinigt analysieren zu können.

In der Einleitung werden zunächst die Konzepte der konvergenten und diskriminanten Validität vorgestellt, Ihre Analyse mittels der Multitrait-Multimethod (MTMM) Matrix (Campbell & Fiske, 1959) ist eine der am weitesten verbreiteten Techniken der Konstruktvalidierung in der Psychologie (siehe etwa Eid, Lischetzke, & Nussbeck, 2006).

Moderne Weiterentwicklungen des ursprünglichen Ansatzes zu CFA-MTMM Modellen erlauben es, die Reliabilität sowie die konvergente und die diskriminante Validität bereinigt um Messfehlereinflüsse zu bestimmen. Darüberhinaus können methodenspezifische Effekte bestimmt und mit anderen Variablen in Beziehung gesetzt werden. Allerdings wurden und werden die meisten Modelle zur Analyse von MTMM Datensätzen für Modelle mit metrischen latenten Variablen entwickelt.

MTMM Modelle für kategoriale Variablen, die die gesamte in einem Datensatz vorliegende Information nutzen, erweisen sich als eine theoretisch sinnvolle Erweiterung zu den bisher vorliegenden Modellen. Sie ermöglichen es, die konvergente und diskriminate Validität auf der Ebene der einzelnen latenten Kategorien zu untersuchen. Das heißt, man kann feststellen, ob bestimmte Kategorien von Konstrukten (z. B. sehr neurotisch zu sein) gut erkannt und somit von verschiedenen Ratern kongruent eingeschätzt werden können und ob andere Kategorien (z. B. nicht neurotisch) nicht genau so gut kongruent eingeschätzt werden können. Die dadurch hervorgerufenen Unterschiede in den latenten Verteilungen werden explizit und kategorienspezifisch in log-linearen Modellen analysiert.

Derzeit liegen jedoch keine Modellformulierungen für MTMM Modelle mit kategorialen latenten Variablen vor. In dieser Arbeit wird diese Lücke geschlossen, es

werden latente Beurteilerübereinstimmungsmodelle und MTMM Modelle für kategoriale latente Variablen für den Spezialfall von Ratern als Methoden (Kenny, 1995) entwickelt.

Zunächst werden die bereits definierten Beurteilerübereinstimmungsmodelle für manifeste Variablen vorgestellt und ihre Bedeutung für die Analyse von Übereinstimmung und mangelnder Übereinstimmung hervorgehoben (Section 2)

Die Erweiterung der log-linearen Modelle zu log-linearen Modellen mit latenten Variablen (z. B. Hagenaars, 1990, 1993) ermöglicht es, die Übereinstimmung von zwei Ratern bei der Einschätzung eines Konstruktes auf latenter (messfehlerfreier) Ebene zu analysieren (Section 4). Zu diesem Zweck werden die verschiedenen manifesten Beurteilerübereinstimmungsmodelle im log-linearen Modell mit latenten Variablen adaptiert (Section 5). Die Bedeutung der Modellparameter wird im Detail erläutert und der Zusammenhang zur Analyse der konvergenten und diskriminanten Validität hergestellt. Insbesondere werden folgende Koeffizienten definiert: kategorienspezifische Übereinstimmung (category-specific agreement rates), Rater Bias (sensu Agresti, 1992) und die Unterscheidbarkeit (distinguishability) von Kategorien.

In den unterschiedlichen latenten Beurteilerübereinstimmungsmodellen können die Rater entweder konstant höher übereinstimmen, in diesem Fall passt ein Quasi-Unabhängigkeitsmodell II (oder ein Modell mit restringierten Effekten für Zellen auf der Hauptdiagonalen), oder in ihrer Übereinstimmung variieren, was zu einem Quasi-Unabhängigkeitsmodell I, einem Quasi-Symmetry oder einem saturierten Modell führt. Diese Modelle bilden das zugrundeliegende Muster von Übereinstimmung und mangelnder Übereinstimmung ab, jedoch sind ihre log-linearen Parameter nicht in allen Fällen einfach zu interpretieren. Aus diesem Grund bietet es sich an, das Verhältnis der modellimplizierten Übereinstimmung zum Produkt der erwarteten Randsummen zu berechnen (category-specific agreement rate). Dieser Wert gibt an, um welchen Faktor die Übereinstimmung überrepräsentiert ist.

Der method-bias type I Koeffizient gibt an, ob sich verschiedene Rater in den modellimplizierten Randverteilungen unterscheiden. Je stärker der Koeffizient von 1 abweicht, desto stärker ist die Divergenz zwischen den Ratern in der Prävalenzrate für die betreffende Kategorie. D. h. dieser Index gibt an, ob die latenten Klassen, in die die Ratings gruppiert werden, gleich groß sind für die beiden Rater. Mittels dieser Werte lässt sich feststellen, ob Rater eine unterschiedliche “Grundwahrnehmung” von Merkmalsausprägungen haben. Sollten diese Unterschiede zu groß sein, so kann nicht davon ausgegangen werden, dass die Rater das gleiche Merkmal beurteilen und von einer Untersuchung der Beurteilerübereinstimmung sollte abgesehen werden.

Die Validität eines Items zur Messung der latenten Kategorien kann anhand der Zweivariableneffekte zwischen Items und latenter Kategorie bestimmt werden. Zu diesem Zweck können auch die bedingten Antwortwahrscheinlichkeiten oder die Effekt-Parameter herangezogen werden. Liegen für bestimmte Kategorien einer latenten Variablen starke Effekte zu einer bestimmten manifesten Kategorie vor, so kann die manifeste Kategorie als “marker” für die latente Kategorie angesehen werden. Die Validität (bzw. ihre obere Schranke die Reliabilität) für alle Items gemeinsam kann (prinzipiell) mit den mittleren Zuordnungswahrscheinlichkeiten bestimmt werden.

Werden die Rater-Agreement Modelle auf Ratings mit geordneten Kategorien angewandt, so kann eine Überprüfung der theoretisch angenommenen Ordnung der Kategorien vorgenommen werden.

In Section 6, werden zwei saturierte Modelle als allgemeinste Beurteiler-übereinstimmungsmodelle miteinander kombiniert. Die Definition dieses Multitrait-Multirater (MTMR) Modells eröffnet weitere Analysemöglichkeiten für die konvergente und diskriminante Validität, Beurteilerübereinstimmung, Moderatoren von Übereinstimmung und raterspezifischen Effekten. In diesen komplexen Modellen mit Zwei-, Drei- und Viervariableneffekten ist eine detaillierte Analyse von Bedingungen und Konstellationen möglich, die zu erhöhter Übereinstimmung und / oder verringerten Abweichungen im Urteil führen. Die Bedeutung der einzelnen log-linearen Effekte auf die Übereinstimmung auf zwei Konstrukten, nur einem Konstrukt oder abweichende Urteile wird im Detail erläutert.

Raterspezifische Effekte können im MTMR Modell mit mehreren Koeffizienten analysiert werden. Der method-bias type I Koeffizient zeigt an, ob sich die Rater in ihren angenommenen Prävalenzraten unterscheiden. Der method-bias type II Koeffizient zeigt an, ob die Rater die verschiedenen Kategorienkombinationen über Traits hinweg unterschiedlich stark bevorzugen, d. h. ob es eine raterspezifische Sicht in Bezug auf den Zusammenhang von Merkmalen gibt. Das MTMR Modell erlaubt es, diese Effekte auch als bedingte Effekte für bestimmte Kategorienkonstellationen höherer Ordnung zu analysieren.

Die diskriminante Validität kann auf der Ebene von Zweivariableninteraktionen untersucht werden oder in Abhängigkeit von Kategorienkonstellationen höherer Ordnung.

Prinzipiell ist sie hoch, je geringer die Effekte für Zellen abseits der Hauptdiagonalen ausgeprägt sind.

Alle latenten Rater Agreement Modelle und alle MTMR Modelle werden anhand empirischer Anwendungen illustriert. Dabei werden die in Sections 5 und 6 für strukturell unterschiedliche und austauschbare Rater definierten Modelle jeweils an einem Datensatz mit zwei Ratern (Selbst- und Fremdeinschätzung oder zwei Fremdeinschätzungen) angewendet. Dabei zeigt sich, dass die komplexen MTMR Modelle mit Mehrvariableninteraktionen mit den vorliegenden Softwareprogrammen nicht geschätzt werden können.

In Section 7 werden die latenten Rater Agreement Modelle und die MTMR Modelle in Hinblick ihre Analysemöglichkeiten der konvergenten und diskriminanten Validität, der Übereinstimmung, der Unterschiede in den Ratings und der Methodeneffekte diskutiert. Darüberhinaus wird das MTMR Modell in den Kontext des Realistic Accuracy Modells (Funder, 1995) gerückt, welches einen theoretischen Rahmen bietet, mögliche Interaktionen in der latenten Tabelle mit Moderatoreffekten von akkuraten Urteilen (accuracy) zu erklären.

Abschließend werden die Schätzproblematik aufgegriffen und Anforderungen an zu entwickelnde Softwareprogramme und Algorithmen formuliert. Sollten diese vorliegen, könnte die Anwendbarkeit des MTMR Modells an großen Datensätzen überprüft werden.

G.2: Lebenslauf

Aus datenschutzrechtlichen Gründen wurde der Lebenslauf nicht abgedruckt.

G.3: Erklärung

Hiermit versichere ich, dass ich die vorgelegte Arbeit selbständig verfasst habe. Andere als die angegebenen Hilfsmittel habe ich nicht verwendet. Die Arbeit ist in keinem früheren Promotionsverfahren angenommen oder abgelehnt worden.

Berlin, 22.07.2008