POLY TOMOUS FACTOR ANALYTIC MODELS IN
DEVELOPMENTAL RESEARCH

A Thesis in
Human Development and Family Studies

by
Daniel E. Bontempo

© 2006 Daniel E Bontempo

Submitted in Partial Fulfillment
of the Requirements
for the Degree of
Doctor of Philosophy

August 2006
The thesis of Daniel E. Bontempo was reviewed and approved* by the following:

Scott M. Hofer  
Associate Professor of Human Development and Family Studies  
Thesis Adviser  
Chair of Committee

Steven H. Zarit  
Professor of Human Development and Family Studies  
Head of the Department of Human Development and Family Studies

Andrea M. Piccinin  
Senior Research Associate, Department of Human Development and Family Studies

Clancy B. Blair  
Associate Professor of Human Development and Family Studies

Jon F. Nussbaum  
Professor of Communication Arts and Sciences

*Signatures are on file in the Graduate School.
Abstract

The relations among constructs in our analytical models of human development correspond to substantive theories, however it is often forgotten that there is a theory of measurement that connects our data to the constructs in our models. This work examines the use of factor analytic measurement models for test instruments with polytomous item response formats. Specific attention is given to issues of factorial invariance, the theorized causal direction between indicators and constructs, and whether the intended use of the test if for diagnostic screening or for individual differences research. When clinical checklists, which often have dichotomous (yes/no) or polytomous (never, sometimes, always) response formats are used in individual differences research, measurement issues need to be reviewed because decisions made during the development of a diagnostic instrument often differ from decisions made in the course of developing an instrument intended to quantify inter-individual and intra-individual differences. In addition, factorial invariance procedures for constructs with categorical (i.e., dichotomous or polytomous) indicators, and the associated confirmatory factor analysis (CFA) model are examined. These goals are advanced through the application of polytomous CFA modeling of two checklist instruments: The Developmental Behavior Checklist, and the short form of the Eysenck Personality Questionnaire. First, a general review is undertaken of the factorial invariance literature, causal directions in factor analytic models, and recent extensions that permit factor analytic models of polytomous test items. Next the factor structure of the Developmental Behavior Checklist is examined using polytomous CFA. Issues of causal direction, factorially complex items, and simple structure are discussed. Finally, polytomous CFA models are used to investigate the factorial invariance of the Eysenck Personality Questionnaire across gender and late-adult age groups. In polytomous factor models a latent continuous response variable underlying each observed categorical indicator is connected to the observed indicators by calculating the cutpoint (or threshold) needed to produce the observed response frequencies. Since these thresholds are all that connects the actual data to the factor model, it is essential that investigations of invariance test the threshold parameters in addition to the usual test of loadings and item-specific variances. However, the additional measurement model parameters in polytomous CFA presents issues for the specification and identification of an appropriate baseline model for nested invariance tests. This problem is addressed by using two interdependent nested sequences. One sequence permits testing of measurement thresholds and loadings, while assuming invariance of item-specific variance, and the other sequence permits testing of the item-specific variances by assuming threshold invariance.
TABLE OF CONTENTS

List of Tables ........................................................................................................... v
List of Figures .......................................................................................................... vi
Acknowledgements ................................................................................................. vii

Chapter 1. INTRODUCTION ..................................................................................... 1
   A Note on Format .................................................................................................... 14

Chapter 2. ASSESSING FACTORIAL INVARIANCE IN CROSS-SECTIONAL AND
   LONGITUDINAL STUDIES .................................................................................. 23

Chapter 3. MEASUREMENT AND METHODOLOGICAL CONSIDERATIONS FOR
   THE DEVELOPMENTAL BEHAVIOR CHECKLIST IN INDIVIDUAL DIFFERENCES RESEARCH
   .................................................................................................................................. 90
   Appendix I - Manifest Change vs. True Change .................................................. 126

Chapter 4. EVIDENCE FOR AGE AND GENDER FACTORIAL INVARIANCE OF
   THE EYSENCK PERSONALITY QUESTIONNAIRE IN LATER ADULTHOOD
   ................................................................................................................................ 128
   Appendix I – Full Text of EPQR-S Items ............................................................... 165

Chapter 5. AFTERWORD ........................................................................................... 173
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>Satisfaction With Life Scale items and average loadings</td>
<td>26</td>
</tr>
<tr>
<td>3-1</td>
<td>Simple-structure and near-saturated five-factor CFA solutions for the ACAD epidemiological sample</td>
<td>106</td>
</tr>
<tr>
<td>3-2</td>
<td>Classification and mis-classification by quintiles of unit-weighted and factor-weighted scores</td>
<td>115</td>
</tr>
<tr>
<td>3-3</td>
<td>Subscale and factor correlations</td>
<td>116</td>
</tr>
<tr>
<td>3-11</td>
<td>Observed raw scores, true scores and change scores for three individuals under two patterns of true score growth</td>
<td>127</td>
</tr>
<tr>
<td>4-1</td>
<td>Estimated parameters in threshold and uniqueness baselines</td>
<td>142</td>
</tr>
<tr>
<td>4-2</td>
<td>Age, mean EPQR-S scale scores, and reliabilities by gender-age, compared with past studies</td>
<td>148</td>
</tr>
<tr>
<td>4-3</td>
<td>Contrast of P and N means in high vs low Lie-score groups</td>
<td>150</td>
</tr>
<tr>
<td>4-4</td>
<td>EPQR-S Item endorsement rates by age-gender, total, and selected studies</td>
<td>152</td>
</tr>
<tr>
<td>4-5</td>
<td>Cross-tabulation of EPQR-S #06 with EPQR-S #39</td>
<td>156</td>
</tr>
<tr>
<td>4-6</td>
<td>Model fit for baseline and nested models</td>
<td>158</td>
</tr>
<tr>
<td>4-7</td>
<td>Invariance hypothesis tests</td>
<td>159</td>
</tr>
<tr>
<td>4-8</td>
<td>Correlation of Extroversion and Neuroticism across age and gender</td>
<td>160</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Plato’s cave</td>
<td>1</td>
</tr>
<tr>
<td>1-2</td>
<td>Schematic of the factor-analytic measurement model</td>
<td>2</td>
</tr>
<tr>
<td>1-3</td>
<td>Schematic of classical test theory</td>
<td>6</td>
</tr>
<tr>
<td>1-4</td>
<td>Item characteristic curve in the three parameter IRT model</td>
<td>8</td>
</tr>
<tr>
<td>1-5</td>
<td>Parameterization of the polytomous factor-analytic measurement model</td>
<td>9</td>
</tr>
<tr>
<td>1-6</td>
<td>Three illustrations of observed response frequencies and corresponding cut-points</td>
<td>10</td>
</tr>
<tr>
<td>2-1</td>
<td>Common-factor measurement model showing factor means, factor variances, indicator intercepts, and indicator loadings</td>
<td>24</td>
</tr>
<tr>
<td>2-2</td>
<td>Synthesis of invariance terminology and FI test names</td>
<td>36</td>
</tr>
<tr>
<td>2-3</td>
<td>Two-factor, two-group factor analytic structural equation model (FASEM) with intragroup factor covariance</td>
<td>37</td>
</tr>
<tr>
<td>2-4</td>
<td>Factorial invariance and structural invariance terminology</td>
<td>44</td>
</tr>
<tr>
<td>2-5</td>
<td>Multi-occasion invariance model</td>
<td>58</td>
</tr>
<tr>
<td>2-6</td>
<td>Thresholds on latent continuous response variable distributions</td>
<td>72</td>
</tr>
<tr>
<td>2-7</td>
<td>Probability of empty-cell or bias when the population cell probability is .01</td>
<td>76</td>
</tr>
<tr>
<td>2-8</td>
<td>Polychoric correlation estimate and SE when true value is 0.0 and minimum cell frequency is 1, 4, and 10</td>
<td>77</td>
</tr>
<tr>
<td>2-9</td>
<td>Juxtaposition of Millsap's FI procedures for continuous and polytomous indicators</td>
<td>79</td>
</tr>
<tr>
<td>3-1</td>
<td>Comparison of ideal and unfortunate loading patterns</td>
<td>92</td>
</tr>
<tr>
<td>3-2</td>
<td>Comparing unit-weighted and factor-weighted scoring methods using a cutoff value</td>
<td>97</td>
</tr>
<tr>
<td>3-3</td>
<td>Self-Absorbed and Anxiety scoring discrepancies</td>
<td>113</td>
</tr>
<tr>
<td>3-11</td>
<td>A hypothetical nonlinear relationship between unobserved true score and observed scale score</td>
<td>126</td>
</tr>
<tr>
<td>3-12</td>
<td>Observed change magnitude contrasted with true change</td>
<td>127</td>
</tr>
<tr>
<td>4-1</td>
<td>Parameters and constraints in traditional CFA factorial invariance procedures</td>
<td>139</td>
</tr>
<tr>
<td>4-2</td>
<td>Magnitude and pattern of loadings in EFA models</td>
<td>155</td>
</tr>
<tr>
<td>4-3</td>
<td>Inter-factor correlations when using all respondents and only single-siblings</td>
<td>156</td>
</tr>
<tr>
<td>4-4</td>
<td>Sample size and probability of 5 or more events when p = 0.00025</td>
<td>157</td>
</tr>
<tr>
<td>4-6</td>
<td>Comparison of raw scale and factor z-scores for Neuroticism</td>
<td>161</td>
</tr>
<tr>
<td>4-7</td>
<td>Comparison of raw scale and factor z-scores for Extroversion</td>
<td>162</td>
</tr>
<tr>
<td>5-1</td>
<td>Factor-analytic measurement model</td>
<td>174</td>
</tr>
<tr>
<td>5-2</td>
<td>Polytomous factor-analytic model</td>
<td>176</td>
</tr>
</tbody>
</table>
Acknowledgements

This dissertation was a process aided along the way by mentors, colleagues, and friends. Special appreciation is reserved for my advisor and committee chair Dr. Scott Hofer, as well as the other members of my doctoral committee. Much of this work was supported by a National Institute of Aging Pre-Doctoral Training Grant (#5 T32 AG00048-28), and this is gratefully acknowledged. Acknowledgement is also given to the owners or facilitators of the datasets used in chapters two and four: Colleagues with the Australian Child to Adult Developmental study – especially Stewart Einfeld, Bruce Tonge, and Kylie Gray; and Michael Stallings at the Boulder Institute for Behavioral Genetics. Special thanks are also due Andrew MacKinnon who provided extensive feedback on drafts of chapter three, and to Athony Ong who was the corresponding editor that shepherded a version of chapter one into the Oxford Handbook of Methods in Positive Psychology. I would also like to acknowledge and thank David Wagstaff who offered many late-night consultations on a myriad of statistical points.
Chapter 1: Introduction

Measurement is an essential activity that quantifies constructs used in our theoretical and statistical models. The four manuscripts comprising the body of this dissertation focus on the application of polytomous factor-analytic measurement models. As a context, this introduction offers a broader view of measurement and introduces general measurement concepts and issues with traditional factor-analytic models that will not be directly treated, but which are relevant to applications of polytomous factor-analytic models, and will appear throughout the four manuscripts.

Measurement theories are important to social science research because they define the nature of latent constructs associated with the observed scores and articulate assumptions that permit estimation of latent psychometric properties from observed data. Theories relating constructs in our models receive a lot of attention, but often overlooked is that theory is also needed to connect our data to these constructs. Constructs are in essence abstract ideas; they are the unobservable *true forms* which cast the shadows seen on the wall of Plato’s cave. (See Figure 1). The constructs in our models are at best empirical abstractions that approximate the true form; they are proxies based on empirical data (e.g., items, indicators, test questions, measurements, responses) that we can observe. The simplest measurement model might be called *res ipsa loquitur* or the thing speaks for itself, where the observation is taken to be a sufficient proxy for the abstract concept, and any error (distortion) is ignorable for the researcher’s purposes. More complex measurement
models associate observations with latent variables that are held to be better empirical abstractions of the true construct. For example, the factor-analytic (or common-factor) measurement model conceptualizes an underlying error-free latent variable that causes the pattern of observed responses (denoted by the rectangles in Figure 2), and the latent variable (denoted by the circle in Figure 2) is then assumed to be a perfect proxy for the true form of the associated construct. In this sense, at the end of any measurement model is the leap of faith that an empirical construct is a sufficient proxy for the true form – at least sufficiently so for the purpose at hand. Thus, every psychometric measure draws on some theory of measurement, even if the measurement model is not explicitly stated.

True constructs, empirical abstractions, and the even empirical data used in our models can vary in important ways. A construct may be one-dimensional or multidimensional; it may be broad or specific. Likewise the empirical abstraction and the associated measurement model may or may not model the multidimensional nature of a construct, and the empirical data (e.g., scale items, measurements) used may be fully disaggregated and specific or exhibit some aggregation across trials, situations, or occasions. (For extended review and discussion, see Bagozzi & Edwards, 1998; Edwards, 2001; MacKenzie, Podsakoff, & Jarvis, 2005). The empirical data can be continuous and normally distributed (e.g., grip strength, age, ability) or ordinal (e.g., never, sometimes, always), or dichotomous yes/no items. While traditional FA models assume continuous and normally distributed items, more recent extensions permit the use of polytomous

This dissertation examines procedures and practical issues associated with the application of polytomous FA models. Additional measurement model parameters are needed and there are consequential issues for model identification, estimation, and the recommended sequence of analytical steps. Commercial software to estimate confirmatory polytomous FA models (i.e., polytomous CFA) has been available for less than a decade, and advanced features such as the ability to handle missing data and poorly conditioned data have only become commercially available in the last two years. Factorial invariance procedures to check for measurement bias in traditional CFA models are well established (Hofer, Horn, & Eber, 1997; Horn & McArdle, 1992; Meredith, 1993; Widaman & Reise, 1997), although not without controversies of nomenclature and recommended procedures (Vandenberg, 2002; Vandenberg & Lance, 2000). Until very recently (Millsap & Yun-Tein, 2004), procedures for testing polytomous CFA factorial invariance had not been specifically addressed in the peer-reviewed literature. In fact, the didactic treatment of the prerequisite multi-group polytomous CFA framework is also recent (Muthén & Asparouhov, 2002). Also, only the simple case of a single factor model with relatively few items has been discussed. More general factor-analytic issues such as multi-factor models, factorially complex items (i.e., items that associate with more than one factor), or permissible violations of classic factor-analytic assumptions have not received attention. Applications of polytomous CFA models are still infrequent in the literature, and applied reports of factorial invariance tests with polytomous CFA models are virtually nonexistent (Comrey & Lee, 1992).

Different types of constructs and differences in the theoretical assumptions of their measurement models accommodate different types of variables (e.g., ordinal vs. continuous),
require different procedures to estimate the construct, and provide different conceptualizations of noise or bias in measurement. Full review and comparison of major measurement theories is not practical here, but some general discussion of key distinctions and concepts regarding types of constructs and various measurement models is useful to facilitate detailed discussion of polytomous CFA models. General measurement model considerations relevant to a discussion of factor-analytic (FA) models are 1) the characteristics of the constructs being measured, 2) the theorized causal relationship between the observations and the latent construct, and 3) whether the measurement theory provides a methodology to detect measurement bias.

Constructs may be specific (e.g., intentionality to seek treatment in the next two weeks for alcoholism) or general (e.g., satisfaction with life), and may be one-dimensional (e.g., working memory size) or multidimensional (e.g., affective, cognitive, and somatic facets of depression). For multidimensional constructs, the measurement model will fall along a continuum from fully-disaggregated, where each facet has its own sub-model, to fully-aggregated, where each facet is associated with only one observation (Bagozzi & Edwards, 1998; Edwards, 2001). In terms of causality, constructs can be theorized as causing the pattern of observed responses, or the observed responses, taken as a composite, can be held to form the latent construct (Diamantopoulos & Winklhofer, 2001; Edwards & Bagozzi, 2000; Williams, Edwards, & Vandenberg, 2003). For example, satisfaction with work, satisfaction with marriage, satisfaction with health, satisfaction with leisure … etc, might in composite constitute overall satisfaction with life. FA measurement models are only appropriate for latent constructs that are theorized to cause the observations, and its misapplication to other constructs can greatly bias subsequent estimates of relations among constructs (Law & Wong, 1999; MacKenzie et al., 2005).
Perhaps the two most important features of any measurement theory are one the degree to which it facilitates the isolation of random noise and systematic measurement bias, and two any constraints on how the test can be scored and used. For valid inter-individual and intra-individual comparisons in developmental and social-science research, it is crucial that measures display measurement equivalency (a lack of bias across groups or occasions). Investigations of potential bias generally focus on the across-group invariance of defined measurement model parameters and key assumption of the particular measurement theory. Factorial invariance procedures focus on the functional equivalence across groups or occasions of key parameters in the factor analytic model. Of critical importance is that when these parameters are found to be invariant then valid quantitative comparisons can be made using simple sums of responses to scale items (Meredith, 1993).

There are two broad measurement theory divisions. Classical Test Theory (CTT), and its extensions (e.g., Generalizability and Factor Analytic models) fall on one side. Item Response Theory (IRT) and its variants (e.g., Graded Response Theory, Rasch Modeling) fall on the other. CTT related theories assume continuous and normally distributed items, and are considered test-based as opposed to item-focused. CTT assumes continuous and parallel items (i.e., same difficulty, equal variance), decomposes observed item scores into a hypothetical true score (denoted by the circle in Figure 3) and an independent random error component for each item; and CTT provides a framework for widely used test reliability statistics (Traub, 1994). There is no treatment of systematic bias, and the resultant true score estimate may include group-
or occasion-related biases, or other systematic bias such as floor or ceiling effects.

Generalizability Theory (Brennan, 2001; Shavelson & Webb, 1991) and FA models (Bollen, 1989; Comrey & Lee, 1992) relax some CTT assumptions and gain greater resolution of systematic bias. Generalizability Theory relaxes the assumption of parallel items (same difficulty, unequal variances), and focuses on the quantification of multiple sources of measurement error in tests, some of which may be systematic. This may lead to development of an improved measure, or some understanding of the amount of bias that may be in cross-group comparisons, but summary scores are not disattenuated for bias. FA models further relax the assumption of parallel items by allowing different true scores (difficulties) and variances for each item, and regression-like (intercept, slope, residual) parameterization of the relation between the item and the construct. In this case, the residual, also called the uniqueness or specific-factor, is a measurement model parameter representing both random error and systematic variance. With FA theory, investigations of bias focus on these measurement parameters, and when all measurement parameters, including the residual variance unique to each item, are found to be invariant a basis for valid comparisons of manifest summary scores is obtained (Meredith, 1993).

The assumption of continuous data is a crucial point. While some data such as grip strength or age is actually continuous and other data such as word recall might become continuous by virtue of averaging across trials, most questions used on psychometric scales are not continuous. In practice there are a limited number of ordinal response options, and it has been demonstrated that substantial bias can result when the FA model’s assumptions of continuous and normally distributed items are violated (Coenders, Satorra, & Saris, 1997; DiStefano, 2002; Joereskog & Moustaki, 2001). With fewer response options, the continuous
data assumption is more extensively violates, and the greater degree of non-normality. Yes/no checklists might be expected to be particularly problematic – especially when responses are skewed.

The desired use of summary scores is also a crucial point that merits elaboration. Meredith (1993) introduced the concept of the practical test user versus the scientific use of tests to underscore this point: “for the practical user of tests and other measures, strict factorial invariance is essential” (p.542). For Meredith, as he wrote in latter half of the last century, the practical user did not have the computing tools or low-level concerns of the research scientist. The practical user was a consumer of paper-and-pencil tests, for example a personnel director. Scoring was viewed as a simple sum of scale items, and only the final score was of interest. This practical user did not have access to or inclination toward complicated computerized scoring nor was she interested in the probative value of how individual items and/or individual measurement model parameters were performing in different groups, contexts, or occasions. For the scientific user computerized scoring and greater interest in the measurement model parameters was assumed; and depending on the needs of the study, there were potentially less stringent invariance requirements. To begin with, the test for invariant residual variances is obviated if the focus of the study is on the variance of the common factor and its relation to other constructs. Only the measurement parameters needed to insure that the latent variable has the same scale in all comparison groups need to be invariant. Further, if some non-invariance of these parameters is observed, that may be of interest in itself, and further, useful scientific work may still possible under a partial invariance framework (Byrne, Shavelson, & Muthen, 1989; Byrne & Watkins, 2003).
Item-focused theories of measurement such as Item Response Theory (IRT) (Baker, 2001; Embretson & Reise, 2000) are intentionally designed for polytomous items, but do not offer anything to Meredith’s *practical* test user. IRT recasts the concept of true score as individual-ability on a strictly one-dimensional construct, and assume intra-individual and inter-individual responses are then independent (random error) once conditioned on the individual’s ability level (denoted by the $b$ in Figure 4). Given a suitable pool of items with known difficulties, adaptive testing and other efficiencies are achieved. Like test-based theories, biased items are detectable when IRT parameters differ across groups. Unlike FA and other test-based models, IRT models do not assume continuous indicators and in fact require dichotomous or polytomous response. However, scoring requires software that can jointly model the S-shaped item-difficulty and user-ability likelihood functions. The finding that a set of items exhibits no bias across groups or occasions does not permit Meredith’s practical user to make valid comparisons outside of the scoring routines (i.e., use a sum of scale items), although it may be argued that computer technology and/or online test scoring services are more available today than there were in the three decade span in which Meredith developed his ideas (Meredith, 1964, 1993). However, there is still considerable applied research where constructs of interest are measured with simple summary scores, where IRT properties of the scale items have not been determined, where scientific users are behaving

---

**Figure 4.** Item characteristic curve in the three parameter IRT model.

![Item Characteristic Curve](image-url)
like practical users; and so, the need for measurement theories that can validate such use has not been obviated by the advent of IRT.

In this context, extension of the traditional CFA models to include polytomous responses offers the opportunity to properly address the non-normal ordinal nature most psychometric items and preserves the potential to validate the use of simple summary scoring. Polytomous items are addressed by conceptualizing a continuous latent response variable (shown in circles in Figure 5) underlying each observed response. These latent response variables serve as fully compliant indicators for a traditional FA model. Each observed polytomous response is linked to its underlying continuous and normally distributed latent response variable by n-1 threshold parameters, which represent cut-points of the underlying distribution that are needed to produce the observed response frequencies. This is illustrated in Figure 6. This is conceptually similar to conditional probability formulations used in IRT models, and in fact it has shown that polytomous CFA models can be parameterized in ways that yield estimates of IRT measurement parameters (Mehta, in press-a; in press-b). As noted earlier, these thresholds are additional unknown parameters which pose problems for model identification. It is also crucial that these threshold parameters be tested for invariance since they are the only link to the

![Figure 5. Parameterization of the Polytomous Factor-analytic Measurement Model]
actual data. These issues will receive substantial attention throughout the rest of this dissertation. However, polytomous CFA models inherit all of the issues of traditional FA models. This includes the causal relation between the underlying construct and its indicators, the specificity and dimensionality of the construct, and issues related to factorially complex items. These issues will be discussed to the extent that they affect the application at hand.

![Figure 6. Three illustrations of observed response frequencies and corresponding cut-points.](image)

Polytomous CFA models will be applied to existing data from two instruments. The first, the Developmental Behavior Checklist (DBC) (Dekker, Nunn, Einfeld, Tonge, & Koot, 2002; Einfeld & Tonge, 2002), is a 95 item symptom checklist using a tricotomus response format (i.e., never, sometimes, always). A five factor structure was obtained using software capable of exploratory factor analysis of ordinal items: Disruptive, Self-absorbed, Communication Disorder, Anxiety, Social Relations. The second, the Revised Eysenck Personality Questionnaire Short Scale (EPQR-S) (H. J. Eysenck & Eysenck, 1994; S. B. Eysenck, Eysenck, & Barrett, 1985) consists of 48 yes/no items. There are 12 items for each of four subscales: Extroversion (E), Neuroticism (N), Psychoticism (P), and Lie (L, dissimulation and social conformity).

Both of these instruments provide useful opportunities for the illustration of polytomous CFA models, invariance testing procedures, as well as issues of construct characteristics, dimensionality, or factorially complex items. The EPQR-S uses a dichotomous format which presents some challenges that other polytomous formats do not. The underlying E, N, P personality facets are theorized to be causing the observed responses, possibly rooted in stable brain chemistry. The 12 items for each subscale were selected from a larger pool based on their
large factor loadings, factorial simplicity (i.e., substantial loading on only one factor), and internal consistency (S. B. Eysenck et al., 1985). By contrast, the DBC items are a census of problem behaviors and disturbed emotions gleaned from thousands clinical records. Items with as little as 10% variance related to the construct (i.e., a loading of 0.30) were retained, and there is considerable factorial complexity. Finally, although the structure was obtained using factor-analytic methodology, there is no strong theoretical view that the corresponding constructs are causing the item responses. For example, one is said to have a conduct problem because of the multiple Disruptive Scale items affirmed, rather than one affirmed multiple Disruptive Scale items because of an underlying conduct disorder. The application of polytomous CFA models with EPQR-S and DBC data will provide many illustrations of both proximal and distal measurement issues.

The first chapter is entirely didactic while the remaining three are empirical reports, although still intended to illustrate procedures and issues. Chapter one extensively reviews the factor-analytic measurement model and traditional invariance procedures with particular attention to controversies and conflicting recommendations found in the literature by Vandenberg and Lance (2000) in their review of the factorial invariance literature. Issues of causal relations between constructs, dimensionality, aggregation, and areas needing further research are also discussed. The focus is traditional factorial invariance and measurement issues; measurement model extensions that permit polytomous indicators and the requisite modifications to factorial invariance procedures are introduced and given limited treatment. The following three empirical reports focus on applications of polytomous CFA models and any issues presented in the actual application.

The manuscript in chapter two reports the fitting of a polytomous CFA model to a
moderately sized DBC dataset, and discusses measurement issues related to factorial complexity. Correspondence of factor scores and manifest summary scores are considered in the face of factorial complexity. Further the correspondence of change in manifest scores with change in factor scores is discussed, and implications for growth curve modeling of the course of DBC subscales is considered.

The empirical reports in both chapters three and four each offer methodological innovation to refine or extend polytomous CFA invariance testing procedures. Chapter three reports tests of polytomous factorial invariance for the DBC subscales across cross-sectional age groups, and illustrates an innovative methodology to explore item-level sources of non-invariance. When invariance testing of an entire scale does not support invariance hypotheses, it is desirable to identify whether all or only some items are the source of the misfit in invariance models. Item level tests introduce additional issues for model identification (Cheung & Rensvold, 1999; Rensvold & Cheung, 2001), and special procedures are necessary to prevent biased items form invalidating hypotheses concerning non-biased items. Chapter four reports invariance tests conducted on the EPQR-S. Dichotomous items are restricted in the degree to which they afford invariance testing of item thresholds. A two-step strategy is developed that permits the fullest possible testing of threshold as well as other polytomous CFA measurement model parameters.

Issues related to construct specificity, single- vs. multi-factor models, factorially complex items, and item-level invariance testing will not be resolved in these chapters, although a short summary and some general recommendations will be provided in a concluding chapter. Rather problematic issues will receive discussion, considerations for choosing the best models and procedures will be articulated, and methodologies suited to the instruments and datasets at hand
will be provided. In addition, findings of invariances or non-invariances for the DBC and EPQR-S will be added to the literature. Perhaps of greatest importance is the idea that there is no single best solution for some issues and that theoretical considerations, the nature of the construct, and the intended purpose of the analyses should be considered in each case.
A Note on Format

In the preface of his 1966 handbook of Multivariate Experimental Psychology, Raymond Cattell described a preface as “a heart-to-heart talk from writer to reader conveying what might seem too personal or human for the austere content of the text.” He asserted that it could be invaluable in helping the reader to focus, follow, and evaluate the text. Despite an already detailed introduction, and with no intent to compare myself to Cattell, I would like to offer a somewhat location-challenged preface to address the structure of the following chapters and the overall coherence of this dissertation.

The originally conceptualized format for this dissertation was three submitted empirical manuscripts, and at every juncture since that moment, that conception has been altered, tweaked, or to some extent abandoned. I see now that I did not fully appreciate the space a traditional dissertation affords for scholarly discussion and extensive review of related topics. By contrast, manuscripts published in the empirical literature require a sharper focus and there is editorial pressure to limit theoretical treatments, liberal use of illustrative figures and plots, and monograph-like synthesis of the existing literature. I think for certain clear cut research questions using well established methodologies the submitted empirical manuscript is an excellent choice. Initially, I had hoped for a more straightforward application of new but clearly outlined polytomous CFA procedures. A large part of my contribution would be the timely placement of applied examples into the literature. As my work progressed it soon became apparent that a clear procedural outline was not available, and that the particular instruments I was using required consideration of broader measurement issues, and in some cases I was to work out the final methodology on my own.
As I investigated, resolved, or dogged many of these issues, the need for more scholarly discussion and didactic content was at odds with the plan to format the dissertation as a series of submitted empirical manuscripts. I recall many discussions with my Scott Hofer concerning subtle points, historical precedents, or outstanding issues; in which he would eventually say, “This is great stuff. It should defiantly be in your dissertation, but this won’t work in the manuscript.” To which I would reply, “But the manuscripts are my dissertation. If it is not in the manuscript, it will not be in my dissertation.” Over time, we settled into the concept of longer more scholarly chapters for the dissertation which would have a one-to-one or one-to-two correspondence to shorter more focused manuscripts submitted for publication. However, the chapters were still construed as separate manuscripts with their own introductions, literature reviews, illustrations, empirical findings, and references. In essence we had come almost full circle to a semblance of a traditional dissertation that is later digested into several publications. However, by this time substantial work had been completed on drafts for three of the four proposed chapters, and complete reversion to a traditional format was deemed impractical. The resulting hybrid format introduces some redundancies across chapters and tolerates shifts in terminology that reflect my developing understanding over the year and a half in which each manuscript took shape. Also, all but the first chapter is organized by instrument and dataset, and not by issues of polytomous CFA models. It is my hope that some chronology and discussion of choices made over the last year and a half will serve to increase the overall coherence.

At the outset the didactic treatment in chapter 1 was not conceptualized as part of the dissertation. My advisor, Scott Hofer, had been invited to contribute a chapter on factorial invariance to a new handbook of methods in positive psychology. He invited me to join him and contribute a short section on the extensions required for polytomous CFA models. I had not yet
formally proposed a dissertation, but I anticipated one or more studies involving polytomous 
CFA models in several datasets available to our lab. Since I was going to be running polytomous 
CFA models, contributing to the chapter seemed like a straightforward thing to do. There were 
some things about factor-analytic models in general and invariance testing procedures that I was 
unsure of, or simply took on the authority of other scholars, but my involvement was going to be 
limited to a section narrowly focused on polytomous extensions.

As work preliminary to my dissertation proposal advanced, and as my contribution to the 
factorial invariance chapter developed, there was a groundswell in the ancillary things I had to 
learn, figure out, or resolve. I had previously received one semester’s funding as a research 
assistant to examine age-based factorial invariance of the Developmental Behavior Checklist 
(DBC), and the problems and successes I encountered suggested a timely opportunity to expand 
this work into a dissertation. At this time no procedures for polytomous CFA invariance testing 
had been published, so I tried approaches developed for traditional CFA. Many models were 
unidentified or failed to converge, and when they did converge, the results did not support 
invariance for any of the five DBC subscales – much to the annoyance of our Australian 
colleagues Stewart Einfeld and Bruce Tonge, publishers of the DBC. However my belief that a 
dissertation could be achieved was fueled by reliance on established procedures for factorial 
invariance tests in traditional CFA models, and by an early draft copy of Roger Millsap’s 
treatment of polytomous CFA invariance which attested that all was possible. Also exciting me 
was the belief that I could find subsets of the DBC items that were invariant – thus pleasing the 
DBC’ authors and possibly advancing the measurement of developmental psychopathology. 
Roger’s final draft, with all problems resolves and appropriate procedures clearly laid out, was 
expected shortly, and I had read about methodologies to find subsets of invariant items via an
exhaustive series of item-level invariance tests (Rensvold & Cheung, 2001). At this point I was also invited to contribute a second short section to the chapter on factorial invariance, this time discussing item-level tests based on the work of Rensvold and Cheung. Things were looking good and the word hubris never entered my mind.

As issues and delays arose, it soon became apparent that this dissertation was going to be more complex and involved; that the procedures were not to be so clear-cut; that some delays were going to be unavoidable, and that I was going to write the whole chapter on factorial invariance as part of the dissertation. Roger’s final draft was delayed for almost two years for technical reasons. The MPlus software we needed to use did not yet support missing polytomous data, and the DBC had quite a bit. Also, I increasingly began to question the suitability of the DBC and its subscale constructs as good factor-analytic exemplars. The fit of basic CFA models (much less multi-group invariance models) was very poor. In my emerging view, a number of DBC subscales seemed to be caused by their indicators rather than the requisite direction of causing their indicators. Also, I could get small models of individual subscales to converge, although with poor fit; but I could not get the larger multi-factor models to converge. I became unsure if multi-factor models were needed, or if the common practice of using a series of single-factor models was sufficient. I undertook a more extensive, slightly obsessive, survey of the literature and encountered Vandenberg and Lance’s extensive review of the factorial invariance literature, which highlighted some of the inconsistent procedural recommendations I had become aware of, and raised new issues I had not yet considered. I developed a strong feeling that any new book chapter on the topic of factorial invariance not only had to discuss the recent advances I was going to write about, but also had to address the inconsistent nomenclature and procedural recommendations that Vandenberg and Lance had highlighted. In a sense they had cast the dye,
and I felt that any new work could not just set forth one procedure without discussing alternatives, or use a particular terminology without trying to disentangle other meanings given to the same terms and/or other terms used to denote the same concepts. Given my now complete immersion in the factorial invariance literature and strong feelings about the onus any new chapter should take on, Scott Hofer suggested I write the whole chapter and include it as a non-empirical manuscript in my dissertation.

While this was a very practical solution that accurately reflected my scholastic efforts, it did challenge the overall coherence of the dissertation in several ways. First, this book chapter was broadly focused on traditional invariance procedures while the empirical thrust of the rest of the dissertation centered on the application of polytomous CFA models. Second, consistent with a volume on positive psychology, the manuscript drew on examples (e.g., study, instruments) of interest to positive psychology. The remaining empirical chapters addressed developmental psychology and personality.

A few final circumstances help shape this dissertation’s present format of expanded-manuscript chapters. After completing my long scholarly treatment of factorial invariance with special attention to historical developments, recent advances, and unresolved issues, the editors pointed out that I was at 200% of the word limit, and ironically suggested that I remove the historical treatment, about half of the figures, the section on polytomous CFA extensions, and most of the discussion of unresolved issues. Taking over the authorship of this chapter and never having undertaken a book chapter before, I had been unaware of any limits. Further, I was now working on my next manuscript and did not have the time to meet the deadline for such an extensive rewrite. Eventually we compromised on a manuscript with fewer figures that was only about 130% over limit, and which still retained some historical treatment, a shortened section of
polytomous CFA, and some limited discussion of unresolved issues. But, I resolved to use the original longer version in my dissertation. Likewise, the manuscript in chapter two was deemed by colleagues to be too long to publish in any high impact journal. While a straightforward report of a confirmatory polytomous CFA model testing the DBC’s published structure did not demand great length, lengthy discussion and several illustrations were needed to cover myriad issues related to the factorial complex items, the questions of causality invoked by the constructs, and the census-like manner in which items were associated with subscales. I worked very hard during June of 2005 with Andrew Mackinnon (from the DBC’s Australian team) to cut 40% of the manuscript. Once again, I decided that a longer version with most of the original plots and illustrations should be used in the dissertation. With these precedents, development of chapters three and four proceeded with the assumption that a shorter empirical version would actually be submitted.
References


Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.


Chapter Two: Assessing Factorial Invariance in Cross-Sectional and Longitudinal Studies

The issue of factorial invariance (FI), or more specifically the sub-set of FI tests that pertain to measurement equivalence, is in essence an issue of construct validity. At the conceptual level, a measure is valid when it accurately operationalizes the construct it purports to measure. The operationalization calibrates manifest indicators to theoretical constructs, which are latent in the sense that they are not directly observed. When a construct is used across multiple groups of individuals or on multiple occasions for the same individuals, the construct’s measurement is invariant (and scores may be quantitatively compared) only when the construct’s operationalization functions equivalently for each group or occasion. This is defined as measurement invariance and multi-group requirements can be mathematically formulated (Meredith, 1993). In practice, invariance of measurements obtained in different groups is often assumed without formal test. This occurs any time scale items are summed or latent factor scores (a loading-weighted sum of scale items) are calculated across groups or occasions without first establishing that the score’s meaning is invariant for each group or at each occasion. For a construct operationalized as a common-factor of a set of manifest indicators, measurement invariance can be demonstrated by testing a sequence of invariance hypotheses focusing on the loadings, intercepts, specific-factors, and structural elements of the common-factor measurement model. Failure to reject these hypotheses demonstrates the validity of subsequent comparisons across groups or occasions.

---

Factor analytic models attempt to explain the covariances of the manifest indicators by using regression weights and intercepts to model each manifest indicator as a linear composite of a single factor accounting for the common variance of the indicators and a specific-factor composed of systematic variance and error unique to each indicator. Figure 1 illustrates the common-factor measurement model. The latent common-factor is represented by the large ellipse, the manifest indicators with their intercepts ($i$) are represented by the four rectangles, the regression-like loadings ($l$) are shown as arrows, and specific-factors (or uniquenesses) are implied by the arrows on the left, but not explicitly drawn. The magnitudes of the loadings are used to interpret the meaning of the construct because the loading is an indication of the strength of the association between the indicator and the construct. The interpretation of a loading is the amount of change observed in the indicator for a one unit change in the common-factor.

The factors, loadings, and intercepts are referred to as model parameters. The model fit, or how successfully the model explains the observed covariances, is denoted by a $\chi^2$ fit statistic. Several additional non-statistical relative fit indices are also used (Bentler, 1990; Cheung & Rensvold, 2002; Hu & Bentler, 1999).

![Figure 1. Common-factor measurement model showing factor means, factor variances, indicator intercepts, and indicator loadings. (Item subscripts omitted).](image-url)
Factorial invariance is subordinate to measurement invariance. Measurement equivalence/invariance requires relationships between theoretical constructs and their observed measures to function equivalently across groups or occasions. The methodology to demonstrate such equivalence must be worked out for distinct measurement models (e.g., classical test theory, item response theory, factor analytic). Each measurement model attempts to represent the relationship of constructs and observed measures using a parsimonious and useful set of parameters (e.g., loadings, intercepts, factor variances). The factor analytic model uses a linear calibration (i.e., intercept and slope) to relate observed measures to latent constructs. Thus, factorial invariance then refers to definitions and methodology to demonstrate measurement equivalence within the factor analytic model. Measurement equivalence/invariance, as the more general and abstract concept, can be thought of as unrestricted while demonstration of FI must approximate measurement invariance under additional assumptions and limitations of the factor analytic model (Meredith, 1993). At a minimum the construct must be factorial in each group in addition to exhibiting across-group equivalence of key parameters.

As noted above, FI procedures apply a sequence of tests to detect if the parameters of the common-factor measurement model are equivalent in each group. These tests are nested because the same parameters are used in each model, and each successive model imposes additional parameter constraints. As seen in Figure 1, we can test for equivalent loadings, intercepts, specific-factor variances, and for equivalence of the common-factor mean and variance. However, equivalence of the common-factor parameters mean and variance is not required for FI. Factor analytic structural equation modeling (FASEM) simultaneously estimates the model parameters in two or more groups and permits testing of parameter equivalence (Joreskog, 1971). Tests involve the statistical comparison of the fit of a model with the parameters freely estimated
in each group to a nested model with the parameters constrained to be equivalent across groups. The sequence of tests begins with no equivalency constraints, and in each subsequent test applies additional constraints to test the loadings, intercepts, and uniquenesses.

For a more concrete illustration of the importance of FI analyses we offer a hypothetical example using the Satisfaction with Life Scale (SWLS) (Diener, Emmons, Larsen, & Griffin, 1985; Pavot & Diener, 1993). The SWLS is a short, five-item instrument designed to measure global cognitive judgments of one's life. Pavot and Diener note that this scale has shown good psychometric properties across numerous independent samples (e.g., unidimensionality, high and uniform factor loadings, temporal stability, language-translation). There has also been some examination of the measurement invariance of the SWLS (Atienza, Balaguer, & Garcia Merita, 2003; Pons, Atienza, Balaguer, & Garcia Merita, 2000; Shevlin, Brunsden, & Miles, 1998). There five items query ideal life conditions or satisfaction with circumstances. Each item uses a 7 point response option ranging from strongly disagree to strongly agree. The actual items and standardized loadings averaged across several studies reviewed by Pavot and Denier (1993) are shown in Table 1.

<table>
<thead>
<tr>
<th>Loading</th>
<th>Item Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.84</td>
<td>In most ways my life is close to my ideal.</td>
</tr>
<tr>
<td>0.81</td>
<td>The conditions of my life are excellent.</td>
</tr>
<tr>
<td>0.81</td>
<td>I am satisfied with my life.</td>
</tr>
<tr>
<td>0.74</td>
<td>So far I have gotten the important things I want in life.</td>
</tr>
<tr>
<td>0.71</td>
<td>If I could live my life over, I would change almost nothing.</td>
</tr>
</tbody>
</table>

If the magnitude of the SWLS item loadings had a substantially different profile in another group (e.g., .11, .71, .72, .91, .62), the latent SWLS construct would be subject to a qualitatively different interpretation, and unambiguous quantitative group comparisons could not
be made. In Table 1, interpretation of the SWLS construct is dominated by an actual-ideal comparison of life circumstances, but other items have similarly large loadings. In the hypothetical alternate profile given above, the interpretation would be dominated by having gotten important things, the other items would have relatively lower importance, and an actual-ideal comparison of life circumstances would effectively not load on the latent construct. By contrast, consider a second hypothetical alternate profile where only the order of the first two items (i.e., close to ideal, and conditions of life) was reversed across two groups. In this case it would be very difficult to argue that the SWLS construct had qualitatively different interpretations in each group. Unfortunately there is currently no well articulated system for describing degrees of non-invariance, or providing confidence intervals for the bias that might result.

Factorial invariance, like the larger enterprise of measurement, has been fairly well developed by decades of work, but is also an ongoing enterprise with unresolved issues and the frequent requirement for researchers to make judgments about its application. For single-construct models with a unidimensional common-factor like the SWLS example above, procedures are fairly clear. For more complex models with multiple common-factors, multi-dimensional constructs, and/or indicators at different levels of aggregation there are unresolved questions and much more need for the researcher’s judgment. Also, invariance requirements in the context of scale development differ from the invariance needs of a single study using a latent framework and ad hoc measures of more holistic constructs. More precisely, each study’s ability to tolerate some degree of misfit differs. Full discussion of topics such as indicator aggregation, construct breadth, and construct dimensionality are beyond the scope of this chapter, but we will
touch on many of these topics in our discussion. A good synthesis of these topics into a single conceptual framework is provided by Bagozzi and Edwards (Bagozzi & Edwards, 1998).

In this chapter, we provide an overview of established methods utilizing factor-analytic structural equation modelling (FASEM) to evaluate FI in cross-sectional and longitudinal studies. The methodology discussed here is based on the theoretical work of Meredith (Meredith, 1993; Meredith & Horn, 2001) and FASEM estimation procedures described by Joreskog (1971) and Sorborm (1974). In addition, Vandenberg and Lance’s (2000) review of the invariance literature has been indispensable for identifying discrepant practices and nomenclature in the applied literature, as well as highlighting aspects of invariance analyses where further work is needed.

Our intention is to provide a straightforward conceptual and procedural introduction for substantive researchers who have some understanding of FASEM, but are less familiar with the measurement equivalency literature. More specifically we have three main goals:

• Provide a clear conceptual understanding of the meaning, importance, and limitations of FI procedures.
• Deliver a clear procedural guide that details the sequence of FI tests and the parameter constraints employed for each successive test.
• Further assist researchers in the application of FI procedures by providing sufficient introduction to circumstances where theory and/or the investigator’s judgement may need to supersede an inflexible application of the full sequence of FI tests.

Achieving these goals will require some discussion of the factor-analytic measurement model and the key parameters that are involved in FI tests. Also, while an unequivocal procedural cookbook will be of interest to many readers, it would be problematic not to address some of the discrepant recommendations and vocabulary in the invariance literature (Vandenberg & Lance,
because failure to do so would hamper readers who want to pursue the FI literature in greater depth. It is also our intent to highlight the important role that the researcher plays in scrutinizing, judging, and interpreting results throughout FI procedures. Based on these goals, the remainder of this chapter is organized into three broad sections. First we begin by clarifying some important factor-analytic concepts and terminology, noting problematic inconsistencies in the FI literature, and adopting a clearly defined terminology. Next we discuss each of the nested FI tests in greater detail. Conceptual issues as well as more detailed instructions to implement each test are discussed. Finally, we address a selection of special topics, recent developments, and unresolved issues pertaining to factorial invariance.

**Distinctions, Definitions, and Conventions**

The purpose of this section is to provide important distinctions and background information to facilitate a more detailed discussion of nested FI tests. First we note that FI procedures are not suited to all constructs. Next we discuss which parameters in FASEM must be invariant for measurement equivalency and distinguish other parameters for which invariance constitutes a substantive hypothesis. Finally we address some of the disorganization Vandenberg and Lance (2000) noted in the invariance literature, and adopt a clear terminology for the remainder of our discussion. Some effort is made to disentangle discrepancies by considering differing motivations for FI testing, the historical development of FI procedures (e.g., mean structures are a more recent addition to FASEM), and noting when differing terminology is synonymous.
The common-factor measurement model.

At the outset, it is important to make a point about latent constructs, causality, and the common-factor measurement model because FI tests only apply to the common-factor measurement model. Many constructs of interest in positive psychology (e.g., optimism, motivation, intelligence, subjective well-being, hope, self-esteem, perseverance) are operationalized as common-factor constructs, but not all constructs use this model. In the common-factor model, the latent construct is conceptualized as causing the observed item responses, and thus the arrows point towards the indicators. (See Figure 1.) In this sense the items reflect the latent construct; and as purported parallel reflectors of the common latent factor, they would have a large degree of covariation. It is this covariation that makes possible exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). In these analyses the known mean, variance, and covariance of the indicators is modeled in terms of the model parameters (i.e., the variance of latent common- and specific-factor, the loading of the items onto the factors, and the item intercepts). Like residuals in regression analyses, the means of the latent specific-factors are fixed at zero and each item’s loading on its specific-factor is fixed to one. Also like regression, the intercept of the indicator denotes how much of the indicator there is when the amount of the latent factor is zero. This is the common-factor measurement model.

The common-factor is called a ‘reflective construct’, and the indicators are designated as ‘effect indicators’. Indicators (scale items) are viewed as a sample taken from a population of parallel reflectors, and the justification for using a small sample of all potential reflectors lies in their conceptualization as parallel (i.e., interchangeable). In contrast, a latent construct could be conceptualized as emerging from the aggregation of a number of formative indicators. For example the construct of life stress overload might be operationalized as acknowledging a
sufficiently large number of common life stressors such as the death of a pet, or the serious illness of a family member. In this case the indicators are conceptualized as causing the latent construct and are not required to have any particular covariation. In other words, you have life stress overload due in part to the death of your pet as opposed to your pet dying because you have life stress overload. Your family member falling ill does not increase the expectation that your pet will die. This type of latent construct is called an emergent construct, and the indicators are designated as formative indicators. [For further discussion of reflective and formative measurement models see Bollen and Lennox’s (1991) discussion of reflective, formative, and mixed constructs. Also see Edwards and Bagozzi’s (2000) discussion of latent constructs and causality. See Diamantopoulos & Winklhofer (2001) for further discussion of formative measurement operationalization.] The important implication of this short discussion of measurement models is that a determination that a measure utilizes the common-factor measurement model must be made prerequisite to considerations of FI. The consequence of incorrect conclusions that may be drawn under the false assumption of a common-factor measurement model are further elaborated by MacCallum and Brown (1993).

It is also important to understand that a measurement model almost certainly departs from reality, which has important implications for the discussion of FI. George Box (1976) pointed out that all models are wrong, but some models are useful. To begin with, models make assumptions (e.g., the indicators are parallel reflections, or the specific-factor variances are uncorrelated) and estimates of model parameters may be biased when assumptions are violated. We will return to this point later when we talk about a baseline model for nested FI tests. Second, a model is an attempt at a parsimonious representation of reality, or at least reality as represented by empirical data. Nothing about the common-factor measurement model presumes the common-factors to be
real entities. This link is made when the researcher adopts a realist philosophy (Borsboom, Mellenbergh, & van Heerden, 2003). While the theoretical role of the common-factor is an underlying truth that is free of measurement error, and as an entity which is there to be reflected by the indicators, its estimate is only a mathematical partitioning of the indicator variance. EFA procedures operate by extracting common-factors until all that remains is a set of uncorrelated specific-factors. CFA procedures partition out the indicators’ common variance, attribute the remaining variance to specific-factors, and given this partitioning, permit an estimated covariance matrix to be statistically compared to an actual sample covariance matrix. In short, the common-factor estimates depend on the particular indicators. With the inclusion of an additional indicator, or by using one less indicator, or by substituting a parallel indicator, the common variance estimate (in any given sample) changes. The assumption we must make is that the change is small and yields no practical difference in the qualitative interpretation of the construct. We will return to these ideas when we consider the baseline model.

**Local independence**

The central factor-analytic assumption of uncorrelated specific-factors is actually problematic in the context of factorial invariance. The condition of uncorrelated specific-factors is called ‘local independence.’ Local in the sense that conditioned on the common-factor(s), which is purported to explain all of the covariance among any set of indicators, the specific-factors should retain no covariance. As noted earlier, local independence is inextricably bound with the estimation of the common-factor. We have already noted that a model is never expected to perfectly match reality, but the problem in a multi-group context is that what is factorial (i.e., meets the local independence assumption) in the population will generally have correlated specific-factors in selected (e.g., gender, age, culture) sub-groups (Meredith, 1993; Meredith &
Horn, 2001). Correlated specific-factors can also result when several indicators share method variance, or pertain to a sub-dimension of a common-factor that is not quite unidimensional. On the one hand we want the construct to be factorial in each group (including meeting local independence assumptions) and to have equivalent parameters across groups; on the other hand, it can be demonstrated using selection theory (Meredith, 1964, 1993) that local independence in each group is unlikely.

In practice this dilemma is resolved by permitting some correlations among selected specific-factors. This is especially true for the longitudinal case where it is important to allow the specific-factor for each indicator to correlate across occasions. In this sense each indicator shares method variance with its replicate at each occasion. To avoid capitalizing on chance, it is desirable that these correlations be specified a priori. In any event it is incumbent on the researcher to discuss and justify any correlations permitted.

**Identification and scaling**

The issue of identification of CFA models is another area where assumptions of the factor-analytic model create difficulties for FI procedures. A CFA model is said to be *identified* when two conditions are met: 1) the model is mathematically identified, and 2) the scale of the factors has been established. Mathematical identification means that there must not be more unknown parameters to be estimated than the number of known moments (i.e., means, variances, covariances) in the observed data. This generally necessitates three or more indicators for each common-factor, although some multiple-construct FASEM models containing factors with only two indicators might be successfully estimated. Scaling requires that the latent factor be given a metric. In CFA models the relationship between loadings and latent factor variance is indeterminate, because the loadings are in part dependent on the units of the factor. Recall that
the interpretation of the loading is the amount of change in the observed indicator for a one unit change in the latent factor. Yet as an unobserved abstraction the latent factor has no absolute units, and a convenient scaling must be imposed. In practice either the factor variance or one of the loadings is fixed to unity. When an arbitrary indicator’s loading (reference item) is fixed to one, the factor’s variance is in the same scale as that indicator, and the other estimated loadings are relative to the fixed loading. Alternatively, when the factor’s variance is fixed to one, the loading is semi-standardized in that it expresses the change in the observed indicator for a one standard deviation change in the latent factor. Choosing different options for scaling will result in numerically different parameter estimates. Choosing which CFA parameter to constrain for identification purposes may yield more or less conveniently interpretable scaling.

In multi-group FASEM, scaling and invariance have a crucial intersection. It has been asserted that only invariant factors (i.e., they are calibrated to their manifest indicators in an equivalent manner), might be unambiguously compared across groups, but now it can be seen that the problem is due to scaling. Unless there are invariant latent factors, scaling differences will prevent any direct comparison. It is desirable to test all parameters for invariance, but identification requires fixing some parameters. Also, the choice of which parameter to fix (use as the reference item) can yield different estimates. This is called the standardization problem (Reise, Widaman, & Pugh, 1993; Rensvold & Cheung, 2001; Vandenberg & Lance, 2000).

**What must be versus what can be invariant.**

There is some confusion in the literature about the number, naming, and sequence of invariance tests required (Vandenberg & Lance, 2000). Different terms are used for the same concept, distinct concepts are referred to with the same term, important tests are sometimes skipped without comment, and frequently tests that extend beyond measurement equivalency are
presented as if they are required. The issues can be phrased as, *What about the factor needs to be the same in each group?* Figure 1 also illustrates what needs to be invariant across groups to establish measurement equivalency. The common-factor is interpreted by the loadings of its indicators, and equivalent loadings are clearly needed; but four distinct aspects of measurement equivalence can be articulated. [A clear and concise development using matrix equations and attention to the presence or absence of group subscripts (indicating constraints) is provided by Widaman and Reise (1997).]

- **Configural** The same indicators are suitable in each group.

- **Metric** The loading profile must define the same common-factor in each group.

- **Scalar** The intercepts need to be the same across groups.

- **Strict** The specific-factors for each indicator should exert a comparable influence across groups.

Meredith (1993) introduces the terms *metric* and *scalar* for the respective invariance hypotheses related to loading and intercepts, and uses the term *strong factorial invariance* when configural, metric, and scalar invariance hypotheses are all retained. He uses *strict factorial invariance* when invariant specific-factors also obtains. A logic of lexical ordering (i.e., weak, strong, strict) has led to the use of the term *weak factorial invariance* when only configural and metric aspects of invariance are obtained, although this term was not used by Meredith. The term *structural invariance* has been used to refer to any additional invariance hypotheses because these tests pertain to the common-factors and their interrelationships and not to the measurement model of each common-factor.

Some confusion in the literature may have resulted from the existence of two different naming schemes. *Weak, strong, strict, and structural* refer to functional levels of invariance and can be related to the level necessary to support the desired use of the measurement (e.g., manifest
composite, comparison of common-factor means). *Configural, metric, scalar, uniquenesses, factor variances, factor covariance, and factor means* refer to each separate invariance hypothesis. There is generally less confusion where Meredith (1964; 1993) defined terms, and more confusion where subsequent scholars were left to supply names. For example, Meredith did not use *weak factorial invariance* and subsequent scholars introduced it to refer to having retained *configural and metric* invariance hypotheses. However, Vandenberg and Lance (2000, p. 12) associate *weak factorial invariance* with having retained only the *configural* invariance hypothesis. Likewise Meredith did not name the hypothesis of invariant specific-factor variances, and Vandenberg and Lance note a variety of names used to describe this test.

Figure 2 provides a synthesis of Meredith’s (1993) terminology and Vandenberg and Lance’s (2000) recommended names for each FI hypothesis. The leftmost column indicates what

<table>
<thead>
<tr>
<th>Intended Purpose</th>
<th>Meredith’s Prerequisites</th>
<th>Vandenberg &amp; Lance’s FI Test Names</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative Comparability</td>
<td></td>
<td>Configural Invariance</td>
<td>Comparable factor composition (pattern of loadings), same simple structure</td>
</tr>
<tr>
<td>Limited Quantative Comparability</td>
<td></td>
<td>Weak Factorial Invariance</td>
<td>Add equality/proportionality constraints on common-factor loadings</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong Factorial Invariance</td>
<td>Add equality constraints on intercepts (Scalar Invariance)</td>
</tr>
<tr>
<td>Unambiguous Multi-group or Multi-occasion Quantitative Comparability</td>
<td>Strict Factorial Invariance (Weak Measurement Invariance)</td>
<td>Metric, Scalar, and Uniqueness Invariance</td>
<td>Add equality constraints on uniquenesses (i.e., specific-factor variances)</td>
</tr>
<tr>
<td></td>
<td>Measurement Invariance</td>
<td><em>(Implied by Weak Measurement Invariance)</em></td>
<td></td>
</tr>
<tr>
<td>Generalizability, Cross-construct Validity</td>
<td>Structural Invariance</td>
<td>Invariant Factor Variances</td>
<td>Constrain common-factor variances to be equal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Invariant Factor Covariances</td>
<td>Constrain common-factor covariances to be equal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Equal Factor Means</td>
<td>Constrain common-factor means to be equal</td>
</tr>
</tbody>
</table>
type of comparisons could be made as each successive hypothesis is retained. The rightmost column specifies the constraints required to frame each invariance hypothesis. The first three invariance hypotheses are configural, metric, and scalar. These constitute Strong FI. The fourth test is for invariant uniquenesses, and together with configural, metric, and scalar constitutes Strict FI. [It is instructive to note that Strict FI denotes what Meredith (Meredith, 1993, p. 530) defines as weak measurement invariance.] The last three tests extend beyond measurement equivalency and pertain to structural invariance.

When models are extended beyond the single measurement model shown in Figure 1 to include multiple common-factors, their respective measurement models, and the structural

![Figure 3. Two-factor, two-group factor analytic structural equation model (FASEM) with intragroup factor covariance.](image-url)
relations among common-factors (see Figure 3), the distinction of what *can be* invariant versus what *must be* invariant is very important. Misconceptions arise because FASEM permits factor-related invariance tests that extend beyond measurement equivalency concerns. Joreskog stated that, “any parameter in the factor analysis models (factor loadings, factor variances, factor covariances, and unique variances) for the different groups may be assigned an arbitrary value or constrained to be equal to some other parameter. … Various hypotheses can be tested by computing several solutions under different specifications. The method is capable of dealing with any degree of invariance” (1971, p. 409). However, invariance tests of factor variances, factor means, covariances, or structural relations pertain to substantive hypotheses, and do not speak to measurement equivalence. Unfortunately, references to *factorial invariance* in the literature often fail to make this distinction. There are situations where structural invariance tests are of substantive interest, and invariance procedures might be used to test these substantive hypotheses. However, factor means, factor variances, and factor covariances are not expected to be invariant across groups under most selection conditions (Meredith & Horn, 2001), and failure to obtain levels of *structural invariance* (SI) should not be seen as problematic. SI hypotheses may provide further evidence of construct validity (i.e., convergent), but not in the sense of measurement equivalence. The rational for testing aspects of structural invariance should always be articulated.

A more concrete example is provided by contrasting two manuscripts that report tests of structural invariance. In an examination of the longitudinal (across three occasions) invariance of a method factor associated with negative worded items on a self-esteem scale, Motl and DiStefano (2002) test whether the covariance of the method factors for the first two occasions is
equivalent to the covariance of the method factors at the last two occasions. Using the concepts of stationarity and stability, they clearly state the rational for the structural invariance tests.

“Our results demonstrated that the method effects exhibited invariance of factor structure, factor loadings, item uniquenesses, factor variances, and factor covariances across three waves separated by 2 years. These results demonstrate that method effects associated with negatively worded items exhibit stationarity and stability across time, similar to measures of other personality traits. Stationarity demonstrates that the same construct is being measured across time (Pitts et al., 1996; Tisak & Meredith, 1990). Stability demonstrates that the relative ordering of individuals on the construct remains constant across time.” (Motl & DiStefano, 2002, p. 571)

Here the term factor structure is used to refer to configural invariance, and the test for scalar invariance is omitted. Most likely the omission of scalar invariance is because the researchers are not interested in factor means and no manifest use of this method factor is anticipated. In contrast, a report of the invariance of a measure of children’s coping (Cheng & Chan, 2003) is more problematic. Structural invariance tests are constructed, but the rational or interpretation is not articulated.

“… analyses showed that factor loadings, factor variances, and factor covariance were invariant across age and gender. On the basis of the factor analyses, the authors created 2 composite scores, representing control-oriented and escape-oriented coping strategies.” (Cheng & Chan, 2003, p. 261)

Here too the test of scalar invariance is omitted, as well as the test of strict invariance. Yet unlike the first example, these authors do anticipate subsequent use of manifest composite scores. This oversight may be mitigated by their implied intention of placing adolescents into a typology of coping styles rather than making quantitative comparisons, but the potential for systematic bias in the composite scores is a serious concern. In their study, the two factor model accounted for less than 50% of the variance in the manifest indicators, so the invariance of specific-factor influence should not be ignored.
A final distinction between what must be and what can be invariant takes into account that recommendations differ based on the investigator’s ultimate goals. Vandenberg and Lance note that several authorities defer to the needs of the study to determine which tests are necessary, and they assert, “like all statistical analyses, the aims and goals of the study should determine which specific ME/I tests are undertaken and the ordering in which they are undertaken” (2000, p. 18). [Note ME/I is their term for factorial invariance.] The most important distinction is between scale development efforts, where subsequent use of manifest composites is anticipated, and comparisons of common-factors in a FASEM framework.

In the course of scale development, the tests of scalar invariance and invariant uniquenesses are essential because it must be anticipated that the scale (set of indicators) will subsequently be used at the manifest level. In FASEM contexts, non-equivalent specific-factors (and in some circumstances non-equivalent intercepts) may be less important because they are modeled, and do not alter the common-factor. The FI literature has sometimes denoted these hypotheses as of less importance. However, when the scale is summed to create a manifest indicator, systematic bias is introduced if Strict FI has not been demonstrated. This is an important distinction between invariance concerns in FASEM and the manifest use of measurement scales. Classical test theory (Lord & Novick, 1968) views items as composed of true score and error components. Under aggregation the error components cancel each other and the composite is expected to be more reliable and valid than any single item. Unlike error, differences in intercepts and specific-factors are systematic and thus aggregate. The composite may then be less reliable or valid than any single item. It is very important that scale developers, who should anticipate the manifest use of their scale, keep this point in mind when they encounter any claims that tests of scalar invariance or invariant uniquenesses are of lesser
importance. Meredith (1993) introduced the concept of the *practical test user* versus the *scientific use of tests* to underscore this point: “for the practical user of tests and other measures, strict factorial invariance is essential” (, p.542.)

In contrast, if the study is using an established scale and is concerned only with structural comparisons of common-factors across groups, there is greater flexibility. The test for invariant specific-factor variances may be of little interest. If the goal is to examine correlations (not covariances) and standardized relations to other constructs, the test for scalar invariance could be unnecessary. Vandenberg and Lance note,

“a test of scalar invariance (invariant intercepts) was discussed least frequently, followed by the test of equal factor (latent) means. This also is not surprising because (a) location parameters (intercepts) are often treated as being arbitrary and sample specific, and (b) analysis of covariance and mean structures is a relatively recent development in the structural equation modeling literature” (2000, p. 17).

But these FI tests are easily affected with available SEM packages, and when conducted, the ability to compare covariances and unstandardized relationships becomes available. Also, it is not clear that intercepts should be treated as arbitrary. Finally, if the intention is to use FI to support collapsing across groups in further analyses or studies, then even the tests of structural invariance would be necessary. For example if one group had a greater variance due to more extreme true scores, cases from this group would have disproportionate influence in OLS regression with a collapsed sample.

**Adopting a uniform terminology.**

Lack of a standard nomenclature in the invariance literature is plausibly linked to conflicting invariance recommendations, misconceptions, and inconsistent practices found in the applied literature (Vandenberg & Lance, 2000). As mentioned above, the term *factorial*
invariance is often used to loosely refer to both the measurement and structural equivalencies that can be tested in a FASEM framework. Similarly, in longitudinal research the terms stationarity and stability are routinely used to respectively refer to measurement and structural invariance. Other terminology problems add to the confusion. Vandenberg and Lance note (citations omitted):

> “the nomenclature invoked to describe the various ME/I tests listed in Table 1 varied across sources. For example, the test of configural invariance was referred to as a ‘baseline’ model…, a test of ‘equality of factor structures’ …, or of ‘equal number of factors and factor patterns’…. The test of metric invariance was also referred to variously as a test of ‘invariant factor patterns’ …, ‘equality of scaling units’ …, ‘metric comparability’ …, ‘factorial invariance’ …, ‘factor loading invariance’ …, and ‘full measurement invariance’ … Similarly, the test of invariant uniquenesses was referred to as a test of ‘invariant disturbance covariance structures’ …, ‘invariant error variances’ …, and ‘equality of reliabilities’ … These terminological differences created difficulty in making linkages between methodological approaches proposed or adopted across different sources.” (2000, p. 18).

In addition the terms congeneric, tau-equivalent, and parallel-forms are sometimes used in invariance discussions that draw on Classical Test Theory (Lord & Novick, 1968; Traub, 1994); but there is only a limited alignment with the rest of the FI literature. For example see Pitts, West, and Tein’s (1996) discussion of longitudinal measurement invariance and stability.

Congeneric is synonymous with configural invariance and parallel forms is substantially the same as Strict FI. Tau-equivalence requires all within-group loadings for a construct to be equal and does not have a counterpart in the FI literature. Finally, Widaman and Reise (1997) distinguish metric from non-metric invariance to separate the configural or congeneric level from the all other levels of invariance, but this is unfortunate in that metric is a specific test of invariance that Meredith (Meredith, 1993) names.

Some of the difficulty in developing a standard nomenclature for invariance tests may have originated in the fact that Meredith’s (1993) intention was to formulate the properties of
measurement invariance and corresponding approximations possible in a factor-analytic framework; he did not set out to name a series of FI tests. As discussed earlier and noted in Figure 2, he named two of the four tests related to the measurement model (metric and scalar), and defined two desirable conditions (strong and strict) that accrue when the first three and then all four hypotheses are retained. Besides, there is a non-parallel relationship between the definitions of measurement invariance and factorial invariance. As mentioned earlier measurement invariance is a larger more abstract concept that subsumes factorial invariance. In Meredith’s theoretical and mathematical treatment of measurement, the concepts measurement invariance and factorial invariance are carefully defined, and have distinct meanings. Measurement invariance, and Meredith’s less restrictive concept weak measurement invariance are defined in terms of equivalent (conditioned on the grouping variable) distributions and cumulative densities of non-specific functions that relate a manifest variable to a corresponding latent construct. In contrast, Meredith defines the terms strong factorial invariance and strict factorial invariance in the context of the factor-analytic model, which has an algebraic formulation and specifies locally independent (non- correlated when conditioned on the grouping variable) common- and specific-factors. He asserts that demonstrating Strict FI is almost certainly sufficient to establish weak measurement invariance, although weak measurement invariance can hold without Strict FI. Furthermore, measurement invariance cannot be demonstrated with finality, but Meredith asserts that “In many situations weak measurement invariance would imply measurement invariance”(1993, p. 530). Further, Meredith states that these concepts, “are idealizations ... their validity and existence in the real world of psychological measurement and research can never be finally established in practice” (1993, p. 540). Abstractions can be difficult to treat consistently, and subsequent scholars who wanted to
specify a set of procedures for applied practice had to name tests not specifically named by
Meredith and gloss over the distinctions between *measurement invariance* and *factorial
invariance*.

In the interests of a clear and precise terminology, we will use the term *factorial
invariance* to denote both a desirable measurement property and a set of FASEM procedures;
also we adopt Vandenberg and Lance’s (2000) names for the specific invariance hypotheses.
(See Figure 2). Vandenberg and Lance prefer the term *Measurement Equivalence/Invariance
(ME/I)* over *factorial invariance*, but this is rejected for several reasons. First the FI literature
stretches back over too many decades. Second, ME/I does not sufficiently underscore that these
procedures apply only to factor-analytic measurement models. We further define *factorial
invariance* as denoting only the invariance hypotheses concerning parameters in the
measurement model (i.e., configural, metric, scalar, and invariant uniquenesses). We will take
care to use the term *structural invariance* (SI) to refer to invariance hypotheses about FASEM
parameters that are not part of the measurement model. This terminology is summarized in
Figure 4.

<table>
<thead>
<tr>
<th>Factorial Invariance</th>
<th>Configural Invariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric Invariance</td>
<td>Weak FI</td>
</tr>
<tr>
<td>Scalar Invariance</td>
<td>Strong FI</td>
</tr>
<tr>
<td>Invariant Uniquenesses</td>
<td>Strict FI</td>
</tr>
<tr>
<td>Structural Invariance</td>
<td></td>
</tr>
<tr>
<td>Invariant Factor Variances</td>
<td></td>
</tr>
<tr>
<td>Invariant Factor Covarianes</td>
<td></td>
</tr>
<tr>
<td>Invariant Factor Means</td>
<td></td>
</tr>
</tbody>
</table>
Factorial Invariance Procedures

The goal of this section is to deliver clear procedural information as well as further conceptual understanding of each FI or SI test. We begin with consideration of the data to be analyzed and comments on the order of FI tests. Next we discuss each FI/SI test in detail.

Analysis of FI must be based on raw data or a covariance matrix augmented with means. This follows Meredith’s statement that “Clearly, then, the evaluation of strong or strict factorial invariance requires the modeling of mean vectors as well as dispersion matrices” (1993, p. 534). Comparison of mean structure across groups should be the most anticipated use of developed scales. Analyses based on correlations (where differences in group level and variability have been removed), as is typical (but not necessary) in independent rotation factor analysis (e.g., varimax, promax) and Procrustean rotation, will not establish to a basis for unambiguous interpretation of factors in different groups.

A sequence of models from least constrained to most constrained (i.e., fully invariant) is recommended. Initial work on simultaneous CFA (Joreskog, 1971; Sorborm, 1974) did outline a procedure that began with the fully invariant model and proceeded to models with increasingly relaxed constraints only when fit of the fully invariant model was poor. However, Bentler (2000) has discussed this issue and noted that comparing increasingly constrained models to an unconstrained baseline has the desirable property that each test is independent. When comparisons are instead made to the fully constrained model, then each test is not independent. But in practice, software can generally estimate fully constrained models with less problems than fully unconstrained models because the latter have many more parameters. Ultimately some judgment will be required on the part of the investigator if there are problems estimating the fully unconstrained baseline model we recommend.
Also, an initial omnibus test of equivalent sample covariance matrices across groups is not recommended. Vandenberg and Lance (2000) advocate this initial omnibus test, and stipulate no further invariance tests if this model fits well. However, excellent fit in one part of the model may mask small departures from invariance in other parts of the model. It is also counter intuitive that invariance tests are not made on the actual factor-analytic parameters.

**Configural invariance**

In configural invariance, freely estimated (unconstrained) parameters only require the same indicators to load on the common-factor in each. In a simple measurement model, invariant parameters are not actually being tested; but in models with multiple common-factors invariance of the zero (unspecified) loadings is being tested. In the latter case, this is also a test of simple structure. Good model fit provides evidence that a common-factor with the same indicators exists in each group. Such a model provides evidence for congeneric similarity across groups but does not allow unambiguous quantitative comparisons to be made. This model is important because of its foundational nature. The pairing of indicators and common-factors reflects theory and a priori assumptions of the investigator. This is also the baseline model against which subsequent constrained models will be tested.

We recommend one particular method of identifying the configural model, although there are several options. The factor means and variances must be given a scale, and this must be done in each group. For the specific-factors this is accomplished by fixing the mean to zero and the single loading to one. To identify the common-factor, the mean and variance of the factor could be fixed, or the intercept and loading of one of the indicators (reference item) could be fixed. A third option involves fixing the factor estimate (i.e., mean, variance) in one group and propagating the identification by constraining the corresponding parameter (i.e., intercept,
loading) of one of the indicators (reference item) across groups. These three alternative baseline models are discussed in detail by Reise, Widaman, and Pugh (1993). Each involves the same degrees of freedom, and will yield the same model fit. Only the scaling of the parameters will differ. For reasons not discussed here, the third option (fix mean and variance in the 1st group, constrain one indicator’s intercept and loading to be equivalent across groups) is recommended. This permits a truly nested sequence of models, and follows Reise, Widaman, and Pugh (1993), and Widaman and Reise (1997). Other options, including discrepant choices for the means and variances can be used if they result in a scaling more convenient to the researcher.

Choosing which indicator to use for equality constraints can be done on a theoretical or statistical basis. The model fit is not affected by this choice. Theoretically, it might be convenient to choose an indicator which has the highest face validity with respect to the proposed latent construct. On a more statistical basis, an indicator that had the largest loading in prior EFA studies could be chosen. If EFA is conducted within each group, an indicator which had a relatively large loading but which also showed the most consistent loading across each EFA would be desirable because this suggests that this indicator is invariant. The choice does not matter at this point, but a well behaved indicator may facilitate subsequent steps.

To summarize the recommendation for constraints on the configural model:

1. The same model (i.e., the same indicators) is specified in each group.
2. The variance of the common-factor is fixed to one in the 1st group.
3. The mean of the common-factor is fixed to zero in the 1st group.
4. The intercept and loading of a suitable indicator is constrained to be equivalent across groups.
5. All other non-fixed parameters are freely estimated.

Although our recommended baseline model scales the mean and variance of each common-factor in terms of the mean and variance of the common-factor in the 1st group,
interpretation of any differences is still ambiguous because the ratio of these differences may change if a different scaling were used. For example if the common-factor variance in the second group is twice the variance in the first group when the 1st indicator is used for identification, it may not be twice as large in the scaling that results if the 2nd indicator is used instead. For further discussion see Widaman and Reise (1997). To preserve the ratio of factor differences, both the intercept and loading parameters will need to be invariant.

At this point, it is useful to consider different meanings for configural invariance and baseline models. The two terms are generally used synonymously, and they are for FASEM models with a single unidimensional construct. The import of baseline is that it is the model with freely estimated parameters against which subsequent nested models with constrained parameters will be compared. For simple models the implication of configural (when a good fit is indicated) is only that the construct (and the given indicators) is factorial in each group. In the cases of multidimensional constructs or FASEM with multiple common-factors, some thought should be given to univariate versus multivariate FI testing. In these more encompassing models, baseline retains the same meaning, but configural has much greater import. Not only is the suitability of the indicators for each common-factor tested, but in addition simple structure is tested. The additional meaning is that the indicators for each common-factor have no cross-loadings on the other common-factors. This can be problematic because factors are sensitive to their context. Each common-factor varies (hopefully slightly) over alternate sets of indicators, but also each indicator’s cross-loadings (and consequently the definition of each common-factor) vary with the nature of the other common-factors in the model. Vandenberg (2002) discusses an example where a two-construct model of self-esteem and job satisfaction might exhibit simple structure, but a two-construct model with self-esteem and self-efficacy would not exhibit simple structure.
In the later case configural invariance might not be obtained if the associations denoted by the cross-loadings could not be absorbed by the association among the common-factors. In this example we can see that the invariance required by simple structure really pertains to the zero loadings (omitted paths) in the model. This is more demanding than the case with a single unidimensional common-factor for which cross-loadings are meaningless.

Is it better to test the measurement for each common-factor in isolation, or to study measurement properties in the multivariate (i.e. multiple common-factors) context in which subsequent analyses will be conducted? It depends on several considerations. If multivariate tests are conducted and the FI hypothesis is rejected, univariate FI tests will ultimately be required to find whether all or only some constructs lack invariance. Also, multivariate tests could be misleading. Since fit applies to the whole model, it is possible that several perfectly invariant factors could mask problematic non-invariance in another factor. One option is to demonstrate invariance for each common-factor in isolation, and then estimate a multivariate model where measurement parameters are fixed to the previously determined values. There is some question as to the proper degrees of freedom when parameters are fixed in this way. An exhaustive investigation might compare univariate and multivariate analyses to explore any differences in the results obtained. But some common-factors with only two indicators are not identified unless they are estimated in a context that includes other constructs. In addition, multi-trait multi-method models have been developed to properly deal with method variance, and cross-loadings are an integral part of these models. Finally, longitudinal models might want to investigate the lagged effect of specific indicators, and this too requires cross loadings.

Meredith has also discussed simple structure as problematic for invariance. Simple structure is actually an answer to the undesirable tendency of factors to shift definitions in
response to their context. To increase the comparability of factors across studies, a narrower construct defined only by the salient loadings (e.g., above .4) is chosen over the slightly broader construct defined by all loadings. The implicit assumption is that the salient loadings capture the construct’s centroid, and that the small cross-loadings can be ignored in a manner similar to random error. Although for multi-dimensional constructs where it is most reasonable to expect some cross-loadings because all indicators pertain to the same higher order factor, the influence of non-salient loadings should not be treated like error. Meredith initially stated, “simply identified invariant models should be fit first and simple structure specifications introduced subsequently” (1993, p. 542). Later Meredith and Horn made the same assertion, “Let invariance take precedence over meta-theory, such as that of simple structure” (2001, p. 236). They review work on this topic by McArdle and Cattell and conclude that minor loadings cannot be ignored.

“if any nonchance relationships were reflected in even very small pattern coefficients, these relationships would need to be accounted for in the tests for invariance. It seems that in most sets of data in the behavioral sciences small but nonchance factor pattern relationships are the rule even when studies are well designed to indicate a simple structure. The magnitudes of small latent roots of obtained dispersion matrices do not decrease in the manner of error (Horn &r Engstrom, 1979). Given this situation, McArdle and Cattell suggested that investigators should probably seek evidence of factorial invariance by evaluating saturated factor patterns rather than considering only the most simple of simple structures.” (Meredith & Horn, 2001, p. 209)

However it must be noted that this recommendation has been largely ignored. Further many applications benefit from simple structure. Scale development begs simple structure. Studies using existing measures in complex structural models are also facilitated by simple structure. Perhaps only when an investigator is examining relationships among a number of factors that are defined by indicators available in a given dataset (e.g., the investigator selects four survey items to indicate job commitment) does Meredith’s saturated-model advice become more applicable.

Once again, we see that there is no single correct answer and that the purpose of the study, the
need for fine-grained versus more holistic constructs. The degree of imprecision that can be tolerated must also be taken into account.

**Metric invariance (weak factorial invariance)**

This level of invariance involves additionally constraining the loadings to be equivalent in each group while permitting the factor variances and covariances to vary across groups. That the factor loadings are found to be invariant is not to say that they are identical since the factor variances and covariances are free to vary across groups. Rather, we say that the factor loadings in one group are *proportionally equivalent* to corresponding loadings in other groups. In other words, loadings standardized to the common-factor variance would each differ from the corresponding loading in another group by the same proportion – the ratio of the variance in each group. It is essential that the common-factor variances are freely estimated in all but the first group. This condition is what creates a test of *proportionality* when *equality* constraints are imposed on the loadings. As a more parsimonious model, the fit will be poorer than the fit of the configural model. The question though, ‘Is the fit significantly worse?’ If not, metric invariance has been demonstrated.

To summarize the required constraints:

1. Constrain the loading for each indicator to be equivalent across groups.
2. If an alternative baseline identification was used, free the estimates of the common-factor mean and variance in all but the first group

A statistical test of the reduction in fit is effected by taking the difference of the $\chi^2$ fit of each model. The more constrained metric invariance model will gain one degree of freedom for each parameter estimate eliminated. For example, if there were four indicators and three groups, the baseline model would estimate ten loadings (i.e., three free in each of three groups and one
constrained across groups). The metric invariance model would estimate only four loadings because all four would be constrained to be equivalent across all three groups. The difference is six. The $\chi^2$ difference is tested on six degrees of freedom.

Metric invariance permits interpretation of differences in common-factor variances. Because the slope aspect of the linear indicator-factor calibrations now function the same in each group, each common-factor variance has the same metric (units). At this point standardized relationships (beta weights, correlations) between the common-factor and other constructs could be interpreted. Unstandardized regression weights or covariances are still problematic because these depend on the common-factor mean which is still subject to scaling bias.

**Scalar invariance (strong factorial - metric AND scalar)**

Meredith’s Strong FI requires both metric and scalar invariance, and consequently the indicator’s intercepts are now constrained to be equivalent across groups. This requires the model to account for all mean differences in the indicators solely through the common-factor mean. The intercept anchors the linear indicator-factor calibration, because the intercept value denotes how much of the indicator there is when the common-factor is zero. Recall, this is directly analogous to the intercept in OLS regression, and the common-factor is the predictor. (In fact, since the common-factor mean in the first group is fixed to zero, this effectively centers the common-factor, and thus for the first group the intercept is the observed indicator’s mean.) With the additional constraints on the intercepts, the common-factor in each group is on the same scale. The question is, ‘Does this model fit well enough?’ Once again as a more parsimonious model, the fit will be worse. If the fit does not significantly deteriorate, this is evidence that all group differences in observed indicator means can be explained by group differences in the common-factor mean.
To summarize the required constraints:

1. Constrain the loading for each indicator to be equivalent across groups.
2. If not done previously, the common-factor means must be estimated in all but the first group, where the parameter is fixed to zero.

Invariant intercepts are one of the more inconsistent and misunderstood areas in the invariance literature. Inconsistent recommendations about the need to test for scalar invariance may to some extent be explained by observing that many older studies did not model the mean structure and would not have mentioned this step. Also intercepts have often been treated like error and presumed to have random fluctuations from sample to sample. For example:

“This also is not surprising because (a) location parameters (intercepts) are often treated as being arbitrary and sample specific, and (b) analysis of covariance and mean structures is a relatively recent development in the structural equation modeling literature.” (Vandenberg & Lance, 2000, p. 17)

Meredith (1993) was very clear about the need to test for scalar invariance, and that “mean differences in X between selected subpopulations will be conveyed through mean differences in the common factor Z between subpopulations” (p 534). Widaman and Reise also echo this requirement:

“When invariance constraints are imposed on the elements in both the $\Gamma$ and $\Delta$ matrices, group differences in the mean level on the latent variables become identified in the ARF-invariant fashion. Indeed, we may say that strong factorial invariance constraints must be imposed to represent and test group mean differences on the latent variables in any meaningful fashion.” (Widaman & Reise, 1997, p. 294)

Unfortunately, Vandenberg and Lance obscure this point by suggesting that item intercepts are expected to vary and that there may be good reasons not to conduct this test:

“intercept differences may not reflect biases (undesirable) but response threshold differences that might be predicted based on known group differences (desirable), for example, between inexperienced versus highly experienced employees. Thus, whether this invariance test should be undertaken depends greatly on the substantive context underlying the study. For example, assume that after conducting the tests for invariance, substantive reasons will result
in one group (males) being compared with another group (females) where it is hypothesized that one group should have a higher mean on the construct of interest than the other group. Assuming also that the measure is a valid operationalization of the construct and that the hypothesis regarding group differences is true, then the items underlying that measure should also reflect group differences if mean difference tests were conducted on an item-by-item basis. Hence, a test for intercept or scalar invariance (i.e., no differences between groups) is not appropriate because difference in item location parameters would be fully expected.” (2000, p. 38)

This framing misses Meredith’s point that the common-factor mean must be forced to convey group differences in indicator means. Meredith and Horn (2001, p. 218) reiterate “equate intercepts and set the common factor means to zero for one group.” They also provide an applied example and discuss further steps that may be taken if the scalar invariance model does not fit well enough.

**Invariant uniquenesses (strict factorial invariance – metric, scalar, and uniquenesses)**

This test requires the additional constraint of invariant specific-factor variances as well as the constraints for metric and scalar invariance. This model tests the equivalency of specific-factor influences, and may not be necessary when working with common-factors in FASEM contexts. On the other hand if the scale is being validated for subsequent use as a manifest composite, then this test is very important because, unlike random error, specific-factors biases do aggregate. This point is at the heart of Meredith’s (1993) distinction between *practical* and *scientific* use of measures. Less than fully invariant models may serve the scientist by aligning with a substantial theoretical framework, or by generating new knowledge about the constructs involved. For the practical (i.e., manifest) use of measures, strict factorial invariance should be demonstrated to insure individual are treated with fairness and equity.

**The constraints for this test are straightforward:**

1. Constrain the variance of each specific-factor to be equivalent across groups
There are many reasons to expect that the hypothesis of *Strict FI* will be rejected. Meredith’s selection theory (Meredith, 1964, 1993) suggests that specific-factor variances (and covariances) will generally vary across groups. Also, developmental constructs frequently show age-related increases in mean and variance. It is reasonable to expect this to be reflected in both common- and specific-factors. Yet constraining specific-factor variances to be equal across groups and the manner in which specific-factors are identified (i.e., single loading fixed to one) does not permit these differences.

Although not necessary for subsequent FASEM analyses there are a number of reasons to conduct this test anyway. Specific-factors are not error and it can be informative to scrutinize them. If they are large relative to common-factors the issue of swollen specifics and/or poor study design should be considered. If the reliability of observed indicators is of interest, this test as well as a test for invariant common-factor variances will be needed. Finally, the test is easy to conduct with software available today. If strict invariance is not demonstrated, the investigator can then alert Meredith’s practical users that the measure is unsuitable as a manifest indicator.

The issue of swollen specific-factors and the opposite condition of inappropriately swollen common-factors occurs when insufficient or inappropriate indicators are used. This goes back to the point that a factor (whether common or specific) is defined in terms of the company it keeps. Meredith and Horn (2001) offer the example of indicators for a spatial visualization factor that are also associated with the related concept of spatial orientation. Used with other visualization indicators that do not have a spatial orientation component, the specific-factors for the contaminated items would be swollen by the variance associated with spatial orientation. If only similarly contaminated indicators are used, the specific-factors will be much smaller, but the common factor will contain the variance associated with both spatial visualization and spatial orientation. In the first case knowledge about the true structure of spatial abilities may be lost. In the latter case, the interpretation of the common-factor may be broader than the study design
intended. These issues are related to more general discussion of study design and measurement, and go beyond the scope of FI. The reason for this short tangent is to underscore the point that specific-factors are important and tests of specific-factor invariance or other scrutiny should not be lightly dismissed – especially given the small effort required to conduct and report this test.

**Structural invariance**

The following three invariance hypotheses involve equivalence of common-factor means, variances, and covariances in structural models. The associated invariance hypotheses generally do not pertain to measurement equivalence, although they may have measurement implications (e.g., longitudinal stability). A brief comment on each is provided.

**Invariant common-factor variances.** There are two situations when this hypothesis may have measurement implications: reliability, and subsequent regressions. As noted earlier, strict FI does not permit statements about the reliability of the observed measure in each group. Vandenberg (2000) asserts that reliabilities may be inferred from specific-factors only when the common-factor variances are equivalent. Subsequent regressions of the common-factor on other constructs may be biased if due to restriction of range if there are large discrepancies across groups. Since our recommended model identification has fixed the factor variance in the first group, this FI test can be implemented by fixing the variance in other groups to the same value (e.g., one).

**Invariant common-factor covariances.** Situations in which this hypothesis has measurement implications include longitudinal stability and the structure of second-order common-factors. In longitudinal models high covariances among the common factor for each occasion indicates stability in individual differences (i.e., the rank order remains largely the
same). For further discussion and several examples see Pitts, West and Twin (1996). To implement this FI test

The utility of looking at the common-factor covariances versus a model where the first-order factors are used as indicators of a second-order factors in not clear, and the matter deserves further comment. Since indicators are expected to covary, the latter case is a stronger model. Large and invariant covariances across groups would provide evidence that the measurement model for the second-order factor was invariant. However, the longitudinal case is problematic. If the first order factors exhibited correlated rates of change as is the inevitable outcome when the same explanatory variable(s) is driving the changes, this would prevent across-occasion invariance in the measurement of the higher order factor.

**Invariant common-factor means.** This hypothesis is usually substantive and corresponds to a t-test or ANOVA comparison of means. When there are more than two groups, the test is an omnibus test, and follow-up models that released the equality constraint on common-factor means in specific groups would be needed. The advantage over traditional tests is that the common-factors are presumed to be free of measurement error. Since our recommended model identification has fixed the factor mean in the first group, this FI test can be implemented by fixing the mean in other groups to the same value (e.g., zero).

**Multi-group versus multi-occasion**

Several important differences exist when testing for invariance across occasions as opposed to invariance across groups. Figure 5 shows a three occasion model for a single construct with four indicators. At each occasion, the indicators the same indicators are used. Unlike the cross-sectional model in Figure 1, which is replicated in each group, this is a single
group model containing all three occasions. A second difference is that the specific-factor for each indicator is permitted to covary with the specific-factors for the same indicator at each subsequent occasion. This presumes method variance specific to each indicator.

![Figure 5. Multi-occasion model.](image)

When multi-occasion models are used in a context where considerable age-based change is expected, bias may result unless age heterogeneity is small compared to the amount of change expected. This issue relates to selection theory (Meredith, 1964) and has been briefly discussed (Meredith & Horn, 2001, see pp 225-226), but a full and clear treatment is still needed. To illustrate, consider a sample of individuals between 20 and 25 years of age at the first occasion, and successive occasions eight years apart. The expectation is that the construct of interest will develop over the 16 year period studied. However this same developmental gradient would be apparent across the five year age span in the sample. This will violate local independence by inducing positive lagged correlations among different indicators that the common-factors or their covariation is unable to absorb. This can never be fully avoided, and study design must mitigate
this problem by keeping age-bands narrow with respect to the amount of developmental change anticipated over the course of the study.

Special Topics

Special topics that deserve further discussion are partial measurement invariance (PMI), aggregated indicators, relative fit indices, substantive hypotheses that a priori specify conditions where measurement invariance is not expected, and the practical consequences when invariance is not demonstrated. Each of these topics has received some treatment in the literature. Partial measurement invariance (Byrne, Shavelson, & Muthen, 1989) and relative fit indices (Bentler, 1990) may have received the most attention. Both may offer some basis to proceed if necessary levels of invariance are not obtained. Aggregation has received less theoretical attention (Bagozzi & Edwards, 1998; Little, Cunningham, Shahar, & Widaman, 2002), but is often seen in practice. Finally, it is generally understood that we can learn something fundamental about our measurement and conceptualization of constructs when invariance does not obtain; this is the consolation prize when invariance is seen as prerequisite. However, there are interesting substantive hypotheses, especially in developmental or intervention studies, that can be supported by demonstrating that invariance does not obtain. While more questions than answers are provided in this section, these topics lead to increased understanding of FI meanings and limitations, as well as the importance of the investigator’s use of theory and judgment.

Partial Measurement Invariance

Partial measurement invariance (Byrne et al., 1989) is the idea that some invariance (some parameters fixed and others free across groups) is better than abandoning analyses when confronted with lack of invariance. Recommendations about acceptable amounts (how many
parameters) and locations (e.g., loadings, intercepts, specific-factors) of partial measurement invariance (PMI) vary from as few as two (Byrne et al., 1989) to a preponderance of the most salient loadings (Reise et al., 1993). Byrne and colleagues introduced the concept in the context of metric invariance. The argument was that if two or more loadings were invariant then the metric of the common-factor was equivalent across groups, and factor comparisons could still be made. Intercepts were not discussed although ‘mean structures’ was in the article’s title. It might be expected that common-factor means would exhibit the most bias under PMI, and the rational for PMI of intercepts is much less clear. While invariant specific-factor variances are not needed for FASEM analyses, PMI of specific-factors does not eliminate bias for Meredith’s practical test user. Vandenberg and Lance (2000) reviewed fourteen studies whose primary focus was the extension of CFA methodology to test aspects of FI, and found that tests of PMI were the most frequently discussed additional tests, and some investigators tested PMI in conjunction with each test of measurement and structural invariance.

Intuitively, some invariance should yield less bias than no invariance, but this has not been proven, more importantly the improvement is not quantified. Vandenberg (2002) states,

“the structure of research on the issue may be something as simple as working it out algebraically in equation form to show how the matrix of fixed and free factor loadings works through the matrix of latent variables to determine the latent means. On the other hand, the research may be more complex, possibly requiring a simulation approach. Regardless of what form the research stream takes, it will be necessary to show how the means are adjusted for in the presence of partial metric invariance (or, for that matter, partial invariance of any of the other tests subsequent to the metric invariance test)” (Vandenberg, 2002, p. 152)

Aside from how improvement is affected, it is not clear how much improvement is obtained. This is an instance of the larger question, How much bias does lack of invariance introduce? Rejection of an invariance hypothesis means that the discrepancy introduced by invariance
constraints was significant. It does not indicate how much bias is involved. This would be a function of the sample size, the complexity of the model, and the narrow versus holistic nature of the construct. Consider OLS regression where highly significant effects might be found but the magnitude of these effects has little practical consequence. What is needed is knowledge of the threshold where lack of invariance presents significant practical bias for a particular study. What we have though is a statistical test of when factor-analytic parameters are significantly different. Because there are presently no definitive answers, sound theoretical frameworks, understanding the meanings and limitations of invariance tests, and good judgment on the part of the investigator are all crucial. For more detailed discussion see Vandenberg’s (2002) discussion of sensitivity and susceptibility in FI testing.

**Locating misfit within invariance models.** Closely related to PMI is the question of which parameters need to be unconstrained in order to obtain a good fitting PMI model. Inspection of the unconstrained and constrained models is a possible starting place, but standardization and proportionality may make important inequalities difficult to see. The most desirable starting place would be theoretical expectations for particular loadings or intercepts to be problematic, but in practice locating sources of non-invariance is often more exploratory. Most SEM software implements modification indices (Bentler, 1980; Chou & Bentler, 1990) that identify the most poorly fitting parameters. The values of such modification indices denote the estimated drop in the $\chi^2$ value (misfit) of the model that would occur if the restricted parameter in question were freely estimated. It is very important to understand that this kind of respecification may overfit the model to the data and consequently suffer a lack of generalizability (Tomarken & Waller, 2003).

**Relative fit indices in FI/SI tests**
Recent work to develop criteria for the use of relative fit indices in FI tests (Cheung & Rensvold, 2002) is related to issues of sensitivity. The $\chi^2$ and $\Delta \chi^2$ statistics used to assess model fit and model differences are sensitive to sample size and the number of degrees of freedom. To address this issue, relative goodness of fit indices (GFI), which locate the model relative to the worst or best possible fitting models were developed (Bentler, 1990; Hu & Bentler, 1999). Unlike $\chi^2$, no p-value is involved; rather guidelines are offered for cutoff values that denote good and excellent fit. Many of these indices are routinely produced by today’s SEM software packages. Investigators are advised to obtain convergence of both $\chi^2$ and several GFI, and in cases of very large samples are instructed to attend to GFI information when $\chi^2$ statistics may be too sensitive. Until recently however, there were no $\Delta$GFI guidelines. Cheung and Rensvold (2002) used simulation studies to evaluate various GFI indicators and to determine thresholds for $\Delta$GFI. Much more work in this area is needed. This is only one simulation study. However Zelinski and Lewis (2003) provide an interesting re-interpretation of recent studies of cognitive dedifferentiation based on change in the comparative fit index ($\Delta$CFI) versus $\chi^2$. They found a number of studies where $\Delta$CFI would have led to different conclusions about dedifferentiation (invariance) hypotheses.

**Aggregation**

Aggregation refers to using indicators that are a composite (generally an average) of multiple items. This is more common for ability-related constructs like intelligence, where multiple trials might be averaged, or composite scores on several batteries used as individual indicators. For constructs that entail values, attitudes, judgments, or dispositions, item level indicators are generally used (e.g., self-esteem). Disaggregated models generally yield narrower more circumscribed constructs, but the SWLS used as an example at the beginning of this
chapter is a holistic construct with item level indicators. Pro-aggregation arguments draw on conceptions of error in classical test theory, focus on the poorer distributional properties of items, or cite the benefits of models with fewer parameters to estimate. The cons of aggregation include objection to any manipulation between empirical observation and analysis, loss of opportunity to observe the relationships among variables (i.e., cross-loading, correlated specifics), or potential misspecification of constructs – especially their dimensionality. The use of aggregation has been controversial for decades. For example:

“To some, aggregating items to manufacture indicators of constructs is viewed as a dubious practice at best and cheating at its worst. Moreover, the practice of parceling contributes to the oft-whispered reputation of SEM as yielding a “smoke-and-mirrors” distortion of reality. For advocates of parceling, on the other hand, the practice is viewed as one that puts a fine sheen on an otherwise cloudy and therefore difficult to discern picture of reality. In this sense, the use of parcels in SEM is not seen as invoking smoke and mirrors, but rather as providing a carefully polished mirror of reality that really smokes.” (Little et al., 2002, p. 152)

Does aggregation affect invariance testing? Rather than assert a yes/no or good/bad answer, once again the purpose of the study and the judgment of the investigator come into play. Parceling has the least risk of introducing bias when the construct is unidimensional and the domain is narrow. If there are indicator cross-loads, parceling may provide the appearance of configural invariance, but build undesirable variance into the common-factor. Little and colleagues (Little et al., 2002) offer the example of an item with salient loadings on both depression and anxiety. If aggregated with other items and used to indicate depression, some of the anxiety variance enters the common-factor. On the other hand Bagozzi and Edwards (Bagozzi & Edwards, 1998) note that the fully disaggregated model will be the most difficult to fit and the least likely to demonstrate invariance. If the goal of the study is to understand the psychometric properties of measure, fully disaggregated models will be needed. Findings for or
against invariance are highly informative regarding useful theoretical relationships. If the goal of the study is to examine relations among latent variables, and the theoretical model can tolerate some imprecision in locating the centroids of these latent variables, then aggregation should be beneficial. For example, Little and colleagues concluded:

“both proponents and opponents of parceling are correct some of the time, and neither is correct all of the time. If the goal of an investigator is to model effects of a latent variable at a given level of generality (analogous to general intelligence), then appropriate selection of scales or parceling of items can minimize or cancel out the effects of nuisance factors at a lower level of generality (analogous to verbal comprehension, spatial ability, etc.). In such situations, parceling is warranted. If the investigator’s goal is to represent the dimensionality of the measurement space at the level of the individual tests or items, then the minimizing of lower-level effects would tend to obscure precisely the effects that the investigator intends to study. In situations of this type, parceling is contraindicated.” (Little et al., 2002, p. 171)

These indicated and contraindicated situations determined by the investigator’s desired level of generality in the construct assessed are special cases of the larger issues of sensitivity and susceptibility (Vandenberg, 2002), which may in turn be subsumed under the very broad issue of construct validity (Kane, 2001). Concern focuses on tolerance for a shift in meaning of the common-factor. Little and colleagues (Little et al., 2002; Little, Lindenberger, & Nesselroade, 1999) use the term centroid to refer to the desired true meaning of the construct in a conceptual domain space. They discuss how selection and parceling of items can shift the centroid of the common-factor. As with PMI, aggregation forces attention to the issue of how much shift in the centroid can a particular study tolerate.

One related idea is implicit aggregation. When responding to items that reflect values, attitudes, or judgments, subjects implicitly aggregate across occasions and circumstances. This is analogous to aggregating across recollections of state to provide an estimate of trait. Items could be designed to elicit a greater or lesser degree of implicit aggregation. Consider Denier’s (Pavot,
Diener, Colvin, & Sandvik, 1991) conceptualization of the process by which his SWLS items assess global life satisfaction:

“Life satisfaction refers to a judgmental process, in which individuals assess the quality of their lives on the basis of their own unique set of criteria … a conscious cognitive judgment of one's life in which the criteria for judgment are up to the person. … Individuals are also likely to have unique criteria for a good life as well, which in some cases might outweigh the common benchmarks in importance. Furthermore, individuals may have very different standards for success in each of these areas of their lives. Thus, it is necessary to assess an individuals' global judgment of his or her life rather than only his or her satisfaction with specific domains.” (p. 164).

Pavot and Denier report that the SWLS scale appears highly unidimensional with the common-factor accounting for around 66% of the observed variance. It is unclear how concerns about parceling and dimensionality play out when the aggregation is implicit.

**A priori lack of invariance**

Traditional uses of FI seek to establish a basis for concluding that differences on latent variables are true changes on the construct of interest by demonstrating lack of any measurement bias. For multi-occasion invariance, the true-change is interpreted as development, and evidence of measurement bias precludes unambiguous statements about development because the apparent change may be due in whole or part to changes in measurement properties. This is the threat to internal validity called instrumentation. In fact the often invoked image of the rubber yardstick is a particularly apt analogy for the stretching, shrinking, or shifting of units that a lack of metric or scalar invariance implies. In this sense FI analyses are seen as prerequisite to the substantive questions of interest.

It is also important to consider studies where lack of invariance may be desired in order to further substantive hypotheses. While evidence of differential measurement presents an
opportunity to learn something about the construct of interest, this opportunity is often viewed as a consolation prize. The planned analyses cannot be conducted because measures lack invariance, but something can be learned anyway. On the other hand there are often theoretical reasons to expect changes in the factor structure (i.e., number of common-factors, factor intercorrelations) in a developmental framework. For example, theories of cognitive ability have often postulated an early-life differentiation and a late-life dedifferentiation of intelligence. In early-life one general factor differentiates into multiple distinct factors, and in late-life multiple cognitive factors return to a single factor, or in a weaker scenario, intelligence becomes more unified as the intercorrelations (Abad, Colom, Juan Espinosa, & Garcia, 2003; Anstey, Hofer, & Luszcz, 2003; Schmidt & Botwinick, 1989) of cognitive factors increases. In another example, it would not be too difficult to imagine how the structure of self-perception might be perturbed by puberty. The implicit aggregation of actual-ideal comparisons in physical (attractiveness, athleticism) and social (acceptance, dating) domains should reflect physical, social, cognitive, and identity-related changes that accompany puberty. In this case a hypothesis about the development of self-perception might be advanced by the lack of invariance. Theory based a priori predictions are required.

To further explore this idea, consider the SWLS items in Table 1. The fifth item, “If I could live my life over …”, has the lowest loading. It is also reported to have the lowest item-total correlation across several studies (Pavot & Diener, 1993). Pavot and Denier speculate that this may be because this item refers to the past while the others invoke the present. A substantive hypothesis might be that this item and one or more of the others would load on a second factor in elder samples. Perhaps drawing on Erikson’s idea of integrity versus despair, older adults might exhibit a second factor related to paths not taken. FI hypotheses could be used to test this
hypothesis.

**Alpha, beta, gamma change typology.** In the case of cognitively mediated constructs, three change outcomes are possible. The alpha, beta, gamma change (ABG) typology (Golembiewski, 1989; Riordan, 2001) distinguishes true development on the construct from two types of differential measurement. Alpha change is true change on the construct. Beta change is differential calibration of the measure to the construct. This constitutes an instrumentation threat to validity. Gamma change involves a re-conceptualization of the construct as in the hypothetical SWLS example above. Gamma change entails a validity threat from maturation or history. ABG change and invariance hypotheses to exploit it may be especially valuable to positive psychology because many constructs are cognitively mediated and interventions are of interest.

Further work is needed to synthesize ABG and FI literatures. Procedures have been described to explore ABG change in a covariance modeling framework (Millsap & Hartog, 1988; Schmitt, 1982) at both individual and group levels (Schmitt, Pulakos, & Lieblein, 1984). Metric and scalar invariance as well as differences in common-factor variance are linked to beta change. Unlike multi-group studies where factor variances are expected to vary across groups, in multi-occasion studies (especially closely spaced occasions) a change in common-factor variance at least raises the possibility that measurement recalibration has occurred. Gamma change is linked to configural non-invariance and non-invariant common-factor covariances. Millsap and Hartog (1988) describe a two-group (experimental, control) two-occasion (pre-, post-test) model where invariance hypotheses concern the regression coefficient of the post-test on the pre-test to provide additional information about beta change. Unfortunately some recommended ABG procedures run counter to recommended FI procedures. For example earlier work ignored the intercepts (Schaubroeck & Green, 1989), although later applied work did include intercepts.
(Vandenberg & Self, 1993). Also, procedures described by Millsap and Hartog test for invariant covariances before imposing restrictions for metric and scalar invariance. These discrepancies beg further synthesis.

**Recent Advances**

Factorial invariance models with polytomous (ordered categorical) indicators and robust procedures for item-level invariance tests are two recent important developments that greatly extend FI testing procedures. Polytomous indicators are important because many test and survey items use ordinal response options with five or fewer choices. In some cases indicators are a series of yes/no questions. While the best opportunity to study the relationship of measures and constructs is at the fully disaggregated level, limited response options generally entails a serious violation of the assumption of continuous and normally distributed indicators. Differences in parameter constraints and interpretations in polytomous indicator models are reviewed. The second important development, robust item-level invariance procedures, addresses standardization problems that can bias item-level invariance tests. If the invariance hypothesis for a set of items is rejected, it is desirable to know which specific items are the source of the misfit. However, the necessity to choose a reference item is problematic because bias can result unless the indicator chosen happens to be invariant. Procedures developed by Rensvold and Chueng (2001) address this standardization problem, and permit the detection of item-subsets that demonstrate invariance. This procedure and the utility of invariant sub-sets of items is discussed.

**Polytomous Indicators**
Recent software enhancements which relax the assumption of continuous and normally distributed indicators permit binary and ordinal (polytomous) indicators. Many test or survey items use a binary yes/no response option, or offer only limited choices (e.g., never, sometimes, always). The distributions for these indicators, even when symmetrical, can only offer a poor approximation of a normal distribution, and in practice are often very asymmetric (e.g., zero inflated). Traditional factor-analytic techniques assume continuous and normally distributed indicators. Just as with the Pearson Correlation Coefficient, increasingly biased estimates of association (and biased significance tests) result under increasing departures from normality. Biased estimates of association distort the common-factors and non-normality distorts the overall $\chi^2$ fit statistic. In the context of these assumptions, polytomous indicators have poor distributional properties. This was noted in the section on aggregation, and aggregating polytomous indicators has been one method for using polytomous data in FASEM. This may incur misspecification problems as noted in the aggregation section. At a minimum, it precludes the fully disaggregated model which is necessary for detailed study of measurement relations.

Specification, identification, and interpretation of polytomous models does require greater sophistication. Identification is more problematic and requires greater understanding of the measurement-model as well as greater dexterity with the syntax used to specify model parameters and their constraints. It should also be noted that ordinal indicators with as few as four levels and reasonable symmetry (i.e., a hump in the middle) can be successfully modeled as continuous and normally distributed indicators. Enhancements discussed in this section pertain to dichotomous and tricotomous indicators, or polytomous indicators with four or more levels but significant skew. Documentation, apart from highly abstract mathematical treatments, is very thin, and there are limited applied examples in the literature. Full treatment of this topic with an
applied example is beyond the scope of this chapter. However, many constructs of interest in positive psychology are measured at the ordinal level and often exhibit skewed distributions (e.g., attitudes). The discussion and references in this section are provided as a launching pad for investigators who are motivated to use these models.

MPlus currently offers the best solution to both the common-factor and $\chi^2$ bias issues. Other methods used in the past were Brown’s (1984) asymptotic distribution-free (ADF) estimator (available in some packages, e.g., LISREL), which is designed for continuous non-normal data; but this still assumes continuous indicators and also requires very large (> 10000) sample sizes (SmallWaters Corp, undated; University of Texas Academic Computing and Instructional Technology Services, 2000). Modeling polychoric (ordinal-ordinal) and polyserial (ordinal-continuous) correlations (an option available in multiple packages, e.g., LISREL, EQS, MPlus) permits unbiased common-factors. Combining polychoric correlations with the weighted least squares estimator (WLS), a variant of ADF that performs for smaller sample sizes (i.e., 200), is the best approach. Detailed discussion of these procedures, or comparisons of their implementation across software packages, is largely missing from the literature. LISREL utilizes a pre-processor (PRELIS) to compute the polychoric correlation matrix subsequently submitted to LISREL. A four-part series discussing ordinal data, including a multi-group example and some references to invariance, is provided on the LISREL home page (URL: http://www.ssicentral.com/lisrel/mainlis.htm). Likewise, the producer of MPlus provides a web note (B. Muthen & Asparouhov, 2002) discussing multi-group models with polytomous indicators, including some discussion of factorial invariance procedures. They also provide syntax files for the examples used in their web note (URL: http://www.statmodel.com/mplus/examples/webnote.html#catstudy), as well as syntax for other
ordinal examples (URL: http://www.statmodel.com/mplus/examples/categorical.html). The web note also offers a brief comparison of the LISREL and MPlus implementations. It is asserted in the web note that MPlus is more flexible than LISREL in regard to item intercepts because the polychoric correlations and model estimates are computed at the same time, while LISREL requires the PRELIS pre-processor (2002, p.15). Millsap and Tien (2004) also compare LISREL and MPlus implementations.

Muthen and Asparouhov (2002) also illustrate the close correspondence of Item Response Theory (Baker, 2001) and CFA with polytomous indicators. They show that the IRT sensitivity parameter (IRT-a) can be computed from the ratio of the polytomous CFA loading and specific-factor variance, while the IRT difficulty parameter (IRT-b) can be computed from the ratio of polytomous CFA thresholds and loading. In short having the three CFA parameters permits computation of the two IRT parameters.

The MPlus implementation additionally provides several robust variants of the WLS estimator. Pairwise methods used to produce the polychoric correlation matrix can result in a non-positive-definite information matrix which cannot be inverted in the course of WLS estimation. This is most likely to happen when the sample size is small and the indicators are highly skewed (L. K. Muthen, 2004, Jan 16). The WLSM and WLSMV estimators (B. Muthen, du Toit, & Spisic, 1997) do not require the information matrix to be inverted. In simulation studies the WLSMV $\chi^2$ fit statistic has outperformed the WLS $\chi^2$ fit statistic (L. K. Muthen, 2004, Jan 29). Overall, the development of MPlus has been more consciously extended to polytomous variables than is currently the case with most other SEM software.

**Latent response variables and thresholds.** The polychoric correlations used with polytomous indicators present more identification issues because they entail more unknown
parameters in the form of latent response variables and indicator thresholds. A bivariate normal distribution of two latent continuous response variables is conceptualized to underlie the cross tabulation of each pair of polytomous indicators. Where the polytomous indicator has m levels, m-1 thresholds are said to cut the continuous distribution in a manner that produces the observed counts in the cross tabulation. Figure 6 illustrates this idea for two tricotomous items. For each observed tricotomous indicator, two thresholds are used to cut a latent continuous response variable at locations required to produce the marginal frequencies of the cross tabulation. The correlation is then based on these threshold values. For each continuous manifest item there are three unknown model parameters: loading, intercept, and specific variance. For polytomous manifest indicators the mean (or intercept) and variance of the latent response variable as well as the m-1 thresholds are additional unknown parameters.

Consequently, additional parameter constraints are required to estimate a model with additional latent response variables, and the LISREL and MPlus approaches differ. Muthén and Asparouhov (2002, p. 13) assert that as a consequence of using the PRELIS pre-processor, LISREL forces invariant thresholds across groups, and reduces the problem to a conventional continuous-indicator analysis. The advantage is that the approach is easy to understand, but the disadvantages are that invariant thresholds may be a problematic assumption, and in the dichotomous case, group differences in common-factor variance are not possible. Millsap and Tien (2004) work out minimally necessary across-group
threshold constraints to identify the model and permit multi-group model testing. They require the first of the m-1 thresholds to be constrained across all groups, and a second threshold to be constrained for one reference item in each group. Muthén and Asparouhov use a different approach where thresholds and loadings are constrained in a reduced model and tests are conducted against a full model where the threshold and loading for selected items are freed while maintaining model identification by fixing the specific-variance to unity for the selected items (see footnote 3 on p. 11). As demonstrated by Muthén and Asparouhov’s applied example with partial threshold invariance, the MPlus approach is more flexible. However, this is a departure from the hierarchical approach advocated at the beginning of this chapter, and it is not clear that these full and reduced models are nested since some parameters change from free to fixed in the full model.

In practice, the flexibility to free specific thresholds may be of limited consequence. The applied example where Muthén and Asparouhov (2002) illustrate freely estimated thresholds had eight dichotomous indicators, but they free the threshold for only one or two. This means that their baseline model is the fully constrained model, which is less problematic to estimate. Lack of convergence and other estimation problems are much more likely if Millsap and Tien’s (2004) largely unrestricted baseline is used (i.e., only one item with two thresholds constrained across groups, and only the first level constrained for all other indicators). Convergence problems will increase as the skew of the indicators increases. In some of our ongoing work modeling nine tricotomous items, we were unable to estimate Millsap and Tien’s baseline models with partially unconstrained thresholds and unconstrained loadings. To obtain convergence, we had to use a baseline that constrained both thresholds for each item. This may be data dependent and tied to the close relation of thresholds and loadings in polytomous CFA models. So, at best all threshold
parameters are not subjected to invariance testing, and in some cases it is necessary to forgo explicitly testing any thresholds for invariance, although the fit of a baseline model with constrained thresholds does provide some indication of threshold invariance. Alternatively, the fully-constrained baseline approach that subsequently frees thresholds and loadings for some items can be used with the caveats noted above.

However, for dichotomous items, the baseline parameterization suggested by Millsap and Tien (2004) presents a more serious problem. Since they constrain the 1st threshold level for each item to be equivalent across groups, testing for threshold invariance is always precluded. One alternative is suggested by Muthen and Asparouhov’s method of freeing thresholds and loadings for selected items while constraining specific-factor variances. If this is applied to all items, an identified baseline model is realized that has all thresholds and loadings free. The usual hierarchical tests for threshold and loading invariance and be performed, albeit in a context of constrained specific-factor variances. Likewise, tests of specific-factor invariance can be performed, but only under the assumption of invariant thresholds. In essence, this requires the use of two mutually dependant hierarchical sequences. As long as invariance hypotheses are supported in both sequences, there is no problem. However should invariance hypotheses not be supported in either sequence, the validity of tests in the other sequence is called into questions. In other words, should hypotheses of specific-factor invariance subsequently fail to be supported, the implication for any threshold and loading invariance tests conducted under this assumption is not clear. This method needs application and further study, but the thresholds actually connect the observed data to the measurement model, and it is important that they be tested for invariance.
Polychoric correlations and scaling. Another consequence of latent continuous response variables is that like other latent variables, these variables have no units. This results in an analysis based on correlation metric rather than covariance or product moment information. It is necessary to devise a method to recover across group differences in common factor means and variances that are obscured by the standardizations implicit in correlation metric. MPlus accomplishes this by introducing a scaling factor that expresses the across-group variation in loadings, common-factors, and specific-factors (B. Muthen & Asparouhov, 2002, p. 8-9). When the test of invariant uniqueness is not of interest (i.e., the scale is only to be used in FASEM context), scaling factors have little consequence. However changes in common- and specific-factor variance are both expressed in the scaling factor, so a special alternate model parameterization and additional restrictions are needed to test for invariant uniquenesses. [For details, see (B. Muthen & Asparouhov, 2002, p. 9, Section 4.2).] The additional restrictions require all specific-factor variances in the first group to be fixed to unity (or some other convenient constant, e.g., population reliabilities). This means that the only possible test for equivalent specific-factors across groups involves fixing the specific-factors in all groups to the same values used in the first group. This is more restrictive than just requiring specific-factors to be equivalent across groups, but this constitutes what is possible.

The meaning of threshold, loading, and specific-factor invariance is also unclear in relation to these scaling factors. This issue has not been addressed, and proof is needed that invariance hypotheses will not be incorrectly accepted when scaling factors are absorbing any non-invariance. In this light, the ability to calculate the two IRT parameters from the three CFA parameters may take on new meaning. With three knows, two unknowns can be computed, but not the other way around. This suggests that the thresholds, loadings, and specific-factor
Polychoric correlations and problematic cross-tabulations. Another aspect of polychotic correlation correlations is the problem of empty cells in the bivariate cross tabulation. This correlation is difficult to compute unless the marginal frequencies are equal, and approximations which are vulnerable to bias are generally used (Brown & Benedetti, 1977). When all cell frequencies are greater than five, the bias is negligible, however it increases as cell frequencies approach zero. Empty cells on the diagonal are the most problematic, and in the case of dichotomous items, the correlation goes to +/- 1.0. Common-factors actors are estimated based on the covariance of the indicators, so low cell frequency induced bias toward correlations of greater magnitude are a serious issue, because items pairs with empty-cells in their bivariate cross-tabulation may have disproportionate influence in smaller samples where the probability of an empty cell is greater. Brown and Benedetti offer the example of the table (1,9,9,81) where the polychoric correlation is zero, and the table (0,10,10,80) where the polychoric correlation is -1.0. If the population cell probabilities in the first case are correct (i.e., .01, .09, .09, .81) then the chance of obtaining the incorrect correlation of -1.0 (i.e., an empty-cell) using a sample of N=100 is very high (p=.366). However if a sample
size of 500 is used, the probability of an empty-cell falls to $p=0.007$. Of equal importance is the probability of bias due to a cell count less than five. With a sample size of 100, bias is certain ($p=0.996$), but falls dramatically as the sample size is increased.

Figure 7 shows the probabilities of an empty cell or of having a cell count less than 5 for sample sizes between 100 and 1000, assuming the population (i.e., true) cell probability is 0.01. While the chances of a empty cell quickly fall to zero, it should be noted that the probability of bias does not fall as quickly. The effect of bias due to a cell frequency is illustrated in Figure 8. This illustration uses the population probabilities (.01, .09, .09, .81) and a true correlation of 0.0 as given above. Brown and Benedetti (1977) calculated polychoric estimates of -.2481, -.0257, and -.0074 with expected cell frequencies of 1, 4, and 10 respectively (i.e., sample sizes of 100, 400, and 1000). Also shown is the standard error of the estimate, which is largest when the bias is greatest. The bias toward overestimation is clearly visible.

The concept of a true population cell probability is an important concept that underlies a common practice when empty cells are encountered. A cell count of zero is a sample value and the population value, while small, is unlikely to be exactly zero. The accuracy of integer cell frequencies is ½ unit, and the accuracy of the cell probability is therefore $1/(2N)$. It is common practice to make this substitution for zero cell counts. Bias is greatly reduced, but may still be
substantial. For, example, consider the value of negative 0.2481 shown in Figure 8 when the true value was actually 0.0. This is much close than -1.0, but still far off the mark.

In practice, it is important to check the bivariate cross tabulations for empty and low frequency cells. These are most likely to occur when some items have highly skewed distributions (e.g., only 3 people out of 4000 endorse an item). Estimation of the common-factor may be robust to such situations if there are a sufficient number of additional items pairs with adequate cell frequencies and small standard errors for their polychoric correlation estimates. In the case of many empty-cell issues, a decision must be made whether to estimate the factor model as is, after dropping the problematic item, or to obtain more data. Problem with item-pairs where the items load on two different common-factors in a multi-factor model may only lower model fit or slightly inflate the interfactor correlation. When both items are loading on the same factor, the potential to distort the structure of the common-factor needs to be considered.

**Polytomous summary.** Additional considerations for identification and scaling are introduced when polychoric correlations are used, and several modifications of the FI procedures for continuous indicators are required. Sample size needs to be large, especially when items have skewed distributions and low frequency cross-tabulation cell counts is likely. Several other issues are in need of further attention. Discussion of these matters is very limited in the literature, and some issues remain unclear (e.g., how do multi-group identification and scaling issues translate in to multi-occasion models?). Figure 9 juxtaposes Millsap and Tein’s (Millsap & Yun-Tein, 2004) sequence of FI hypotheses for continuous and polytomous indicators. The main differences are some/all invariant thresholds in the baseline model, lack of a separate test for weak factorial invariance, and fixed specific-factors versus constrained. The fully constrained baseline and partial measurement invariance demonstrated by Muthén and Asparouhov (2002) is
an alternative, and might be extended (as described above) to offer tests of thresholds, loadings, and specific-factor variances in two mutually dependant sequences. Further treatment of these issues, additional comparisons across software packages, and more applied examples should make these procedures increasing accessible to applied researchers. Understanding of the issues presented here will facilitate more informed consumption of new developments.

Figure 9. Juxtaposition of Millsap's FI procedures for continuous and polytomous indicator:

<table>
<thead>
<tr>
<th></th>
<th>Continuous</th>
<th>Polytomous</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Baseline</strong></td>
<td>Intercepts, loadings, &amp; uniquenesses are free across groups</td>
<td>Baseline Loadings &amp; uniquenesses are free across groups. Constrained thresholds</td>
</tr>
<tr>
<td><strong>Weak</strong></td>
<td>Constrain loadings</td>
<td>Weak/Strong Constrain loadings</td>
</tr>
<tr>
<td><strong>Strong</strong></td>
<td>Constrain intercepts</td>
<td></td>
</tr>
<tr>
<td><strong>Strict</strong></td>
<td>Constrain uniquenesses</td>
<td>Strict Fix uniquenesses</td>
</tr>
</tbody>
</table>

**Item-level invariance tests**

Recent work by Rensvold and Chung (2001) provides an alternative to partial measurement invariance (PMI) approaches by identifying subsets of items for which invariance does obtain. Which indicators are invariant across all groups? Rensvold and Chung’s *factor-ratio test* (FRT) is a systematic process of testing metric invariance for individual indicators, and combining the results to identify multiple overlapping item subsets which display invariance. The procedure avoids identification related bias that PMI solutions or less systematic approaches to item level FI testing might introduce. (However, the exhaustive number of models, implies that procedure is exploratory). Item subsets may provide a narrower common-factor, and in some circumstance this may be unacceptable. In other circumstances (e.g., growth modeling of several related constructs), a narrower but invariant common-factor might be preferred. Given the issues related to PMI that were previously raised, the *factor-ratio test* merits some explication. We will
briefly review the motivation for FRT, describe the procedure, and discuss unresolved issues.

When an invariance hypothesis is rejected, attention is directed to the item level to find the source of the non-invariance. These efforts might range from studying the degree of discrepancy observed when parameters are free across groups (i.e., in the baseline model) to an examination of modification indices. Subsequent actions might entail estimating new models which drop indicators that previous models and/or theory suggest are problematic, or utilizing a PMI approach where the problematic indicators are retained but not constrained to be invariant across groups. If the purpose is scale development and a measure with optimal psychometric properties is the desired outcome, dropping problematic items would be most likely – especially if a sufficient number of indicators were being retained. Of course, dropping indicators should not be done thoughtlessly because this may not only narrow the construct but possibly shift its centroid in the conceptual domain. It might even be possible to draw on theory to fix the indicator. By contrast, if the investigator’s requirements are limited to FASEM analyses, a PMI approach might be more likely – especially if most of the indicators were invariant.

Unfortunately, item level analyses present several problems. First, the factor-analytic model is a system, and removing any indicator may change the common-factor or other indicators’ specific-factor or loading. When two or more indicators are not invariant they may mask each other, or the investigator may eliminate or free two items when eliminating or freeing only one would have resulted in a the choice of two slightly different systems in which the retained indicator would have displayed invariance. Second, as a result of identification issues, it is not possible to obtain an unambiguous FI test for a single item. Metric invariance hypotheses cannot be tested when common-factor variances are used for identification because of the need to allow proportionality, so one or more of the loadings must be constrained (or fixed) to obtain
model identification. This means that the FI test of any one indicator occurs in the context of the indicator (or indicators in the case of the fully constrained baseline) used for identification. In other words, more than the desired constraint is actually being tested. Cheung and Rensvold (1999) provide an algebraic demonstration of the potential bias that may result.

The factor-ratio test procedure addresses these issues by multiply testing each indicator in the identification context of each of the other indicators. Algebraic support for this ratio hypothesis is provided in Cheung and Rensvold (1999). The test is symmetrical, so only $n(n-1)/2$ tests are required. In this way, non-invariant pairs are identified. Invariant subsets are then formed by taking each combination of indicators that does not include more than one member of a bad pair. The researcher is now able to select among subsets based on size or theoretical concerns (e.g., preferring the second largest subset because it contains the indicator with the highest loading). It also becomes possible to attend to indicators appearing in all or most sets versus indicators found in fewer sets. This procedure has been described in detail (Cheung & Rensvold, 1999; Rensvold & Cheung, 2001). Rensvold and Chueng also describe algorithms that find all invariant subsets given a list of bad pairs, and provide an applied example (2001).

In our ongoing work utilizing the factor-ratio test procedures, we have introduced several refinements to the FRT. First, we adopt the procedure to the baseline model described earlier where common-factor variance is fixed in the first group and one loading is equated across groups. As described, the FRT procedure achieves identification by fixing one loading across groups. This is trivial because the bias stems from identification based on non-invariant indicators and not whether the loading was fixed or equated. Second, we use $n-1$ baselines, while the applied example supplied by Rensvold and Cheung uses only one baseline. Specifically, indicators two through $n$ are tested against a baseline identified by constraining the first indicator.
across groups, indicators one and three through \( n \) are tested against a baseline identified by constraining the second indicator across groups, and so forth.

The FRT procedure should benefit from further study and development. As described, the procedure focuses on \textit{metric} invariance. Specifically the bad pairs (and subsequent good subsets) are determined based on tests of \textit{metric} invariance. It seems plausible that this could be extended to determinations based on both \textit{metric} and \textit{scalar} invariance, but this has not been explicitly examined. Rensvold (under revision) has also extended this work to subsets of groups, asking which groups can be compared using all indicators. This also has longitudinal implications in the sense asking which occasions (or age ranges) is factorial invariance obtained. While contiguous temporal/age ranges are most intuitive, interrupted ranges should also be possible (e.g., several years before and after puberty, but not at the height of pubertal development). The FRT procedure is also very amenable to automation, and such software would make the procedure practical for a wider range of applied researchers.

The factor-ratio test may not be appropriate for invariance testing of models with polytomous indicators. As noted earlier, introduction of a latent continuous response variable forces a correlation metric, and requires the introduction of a scaling factor to capture across group differences in the common-factor. The factor-ratio test depends on true item standard deviations that are not available in correlation metric. Mehta (2006) has expressed concern regarding this issue, as well as the exhaustive number of pair-wise tests. He notes that when the common-factor variance in each group is fixed at 1.0, multiple estimates of the population ratio of reference group to target group common-factor standard deviations are given by the ratio of each item’s factor loading in the reference group to its loading in the target group. Mehta proposes to compute confidence intervals for these ratios and then to select invariant subsets
based on substantial overlap of the distributions. This method shows promise, but awaits further development. The role of the common-factor mean is not clear. Also, it is possible to independently identify the common-factor standard deviation in terms of item loadings and residuals, or in terms of item thresholds and loadings. The implications if these two methods do not agree are not completely clear.

Summary

Factorial invariance and the larger enterprise of measurement are foundational aspects of good science. This is especially true in the social sciences generally, and particularly in fields such as positive psychology where important objects of study are latent constructs whose nature and validity are simultaneously derived from consideration of a set of interrelationships with other latent and observed constructs. Factorial invariance is also essential for developmental studies that implicitly require the comparability of constructs over occasion (e.g., growth modeling). Many of the conflicting proscriptions in the FI literature can be reconciled in historical context (i.e., mean structure introduced more recently), or by considering the investigator’s purpose in terms of level of holism desired in the construct (e.g., issues of aggregation or PMI) and/or a particular study’s ability to tolerate some ambiguity concerning the construct’s centroid (manifest versus FASEM uses and Meredith’s practical user of tests). FI procedures are well established and increasingly facilitated by powerful SEM software packages, and we have carefully described the model parameterization and constraints recommended to test factorial invariance. However, the investigator’s perspicuity and judgment regarding FI and larger measurement issues can not be automated or reduced to the inflexible application of hard rules.
One of the most desirable developments would be theory and methodology to set boundaries on the biases associated with non-invariance. Analogous to moving from p-values to confidence intervals in the more general statistical literature, new methods might replace scenarios that reject invariance hypotheses at a particular probability level, with confidence intervals for the evaluation of substantive effects that become increasingly wider under increasing conditions of non-invariance. This would subsume issues of configural invariance, PMI, aggregation, narrow versus holistic constructs, and sensitivity or susceptibility.

Further treatment of the alpha, beta, gamma change typology (Golembiewski, 1989; Riordan, 2001) should be very useful to positive psychologists. Further integration of ABG and its associated methodologies (Millsap & Hartog, 1988; Schmitt, 1982; Schmitt et al., 1984; Vandenberg & Self, 1993) with established FI procedures should be of particular benefit to positive psychology because of its potential in developmental and/or intervention research. Substantive hypotheses about the effects of development or intervention could be tested with a priori hypotheses of non-invariance at specific locations within the FASEM framework.

Factorial invariance, like measurement, is an ongoing enterprise, and we have endeavored to provide a conceptual map of current limitations and unresolved issues that should allow readers to be informed consumers of subsequent developments in the FI literature. At the same time, FI procedures are currently at a high level of development and should be rigorously applied today, without waiting for future developments.

“Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”

References


Genetic Psychology, 164(3), 261-266.


Muthen, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika,* (You can get a copy of this paper by emailing bmuthen@ucla.edu. It was never revised and resubmitted).


Rensvold, R. B. (under revision). Metric equivalence/invariance across multiple groups: Comparing apples with apples with apples etc. *Organizational Research Methods*.


Vandenberg, R. J. (2002). Toward a further understanding of an improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.


Chapter 2: Measurement and Methodological Considerations for the Developmental Behavior Checklist in Individual Differences Research

The Developmental Behavior Checklist (DBC) (Dekker, Nunn, Einfeld, Tonge, & Koot, 2002; Einfeld & Tonge, 1995) is a 95 item clinical screening checklist designed to assess the extent of behavioral and emotional disturbance (BED) and several factor-analytically derived BED subscales (e.g., Self-Absorbed, disruptive) in populations with intellectual deficit (ID). The DBC is in broad international use, and is recommended in reviews of psychopathology (Aman et al., 2004; Wallander, Dekker, & Koot, 2003). The DBC currently has a primary caregiver version (DBC-P), a teacher version (DBC-T), and a recently developed version for use with adults (DBC-A) (Mohr, 2003). The revised five-factor structure of the DBC (Dekker, Nunn et al., 2002) was extracted using principal components analysis (PCA) in a large international (Dutch-Australian) sample of young people with a full range of intellectual deficit. The DBC’s Total Behavior Problem Score (TBPS) constitutes an index of individual and family burden, while the five subscale scores (Disruptive, Self-Absorbed, Communication Disturbance, Anxiety, and Social Relations) offer greater resolution over clinically relevant dimensions of psychopathology, and provide a basis for studying the natural history of BED in ID populations.

Use of the DBC to study individual trajectories of BED invites reconsideration of the DBC’s measurement properties. The DBC was designed as a diagnostic screening instrument, and in comparing the task of diagnostic classification as case/non-case to the task of quantifying intra-individual and inter-individual differences, two qualitatively different endeavors are revealed. In individual differences research the primary goals involve both between-person comparisons and studying within-person change; and decisions about number of items loading on a factor to retain in a scale (i.e., cutoff criteria) and the best scoring procedure for subscale
totals may not provide the most sensitive and reliable instrument for quantitative comparisons if these decisions have been optimized for screening purposes. Scale development for screening purposes focuses on building a broad item bank and retaining items. In contrast, developing scales for individual differences research often focuses on eliminating factorially complex items and retaining only sets of items with high and homogeneous loadings. This avoids inflated correlations among unit-weighted subscale scores, and also provides greater qualitative (i.e., construct) correspondence between the unit-weighted score and latent factor estimates.

The best decisions reflect both the intended use of the measurement and the measurement properties of the scale’s items. For example, unit-weighted (i.e., simple sum) scoring is robust to sample-dependent fluctuation in factor loadings, however factor-estimated scores (or other work with latent factors) may be more sensitive to conditions that markedly alter factor loadings and consequently shift the centroid of the construct in the conceptual space. Likewise retaining many items, including those with modest loadings, will increase the screening sensitivity to a wide array of problem behaviors and ensure a high negative predictive ability. However, for quantitative comparisons, the utility of including items with moderately low factor loadings (e.g., 0.30) in unit-weighted scoring procedures is less clear because of potential problems introduced by non-negligible secondary loadings for any items which are factorially complex. The import of many small secondary loadings has been discussed in the context of the frequently observed poor fit in CFA studies of EFA derived structures (Borkenau & Ostendorf, 1990; Van Prooijen & Van Der Kloot, 2001). These researchers have questioned the applicability of CFA with multidimensional instruments comprised of broad high-level factors, but CFA provides the only direct test of a structure’s generalizability, and methods to test hypotheses of factorial invariance require a CFA framework.
Measurement concerns can be broadly grouped as either structural or scoring. Structural concerns focus on the generalizability of the structure, and whether items continue to associate with the same factor in successive studies and generally do not focus on the specific magnitudes of factor loadings. However, substantial changes in the magnitude of factor loadings can be important if they constitute a large shift of the factor in conceptual space. Scoring concerns focus on the distributions of unit-weighted and factor-estimated scores and the extent to which they are discrepant.

**Structure**

It is important to note that in general CFA practice, the item loadings are still freely estimated just as they are in EFA, and the solution must be inspected to determine if an equivalent (congeneric) factor is being obtained. The fit of a CFA model is an indication of the cost of imposing a simple structure where items are assigned to certain factors and, more importantly, where other items are prevented from loading on that factor. The constraints define each factor by specifying which items do not load on the factor. In this sense, CFA model fit does not directly indicate congruence of factors in CFA models with similar factors extracted in prior PCA models.

![Figure 1. Comparison of ideal and unfortunate loading patterns.](image-url)
When the constrained loadings (or cross-loadings) in a simple structure are not truly negligible, imposing simple structure can distort CFA solutions. Factor loading estimates may shift in the constrained simple-structure model because only the explanatory ability of the relatively fewer unconstrained loadings is being maximized. As the factor moves in conceptual space, there is often a complex shuffling of the salient item loadings which can obscure factor congruence or non-congruence. The conceptual shift is most likely to happen in the context of factorially complex items and least likely to happen in more ideal circumstances where each item has a large association with only one factor. In ideal circumstances salient loadings are high, well above the usual cutoff (0.30), and fairly homogeneous; likewise, nonsalient loadings are low or near zero (i.e., in the hyperplane). Ideal and unfortunate loading patterns are illustrated in Figure 1. The imposition of simple structure is not likely to alter the solution or increase misfit because the nonsalient loadings taken together only explain a small amount of item variance. With oblique rotation, changes in factor loadings are further amplified because the covariance among items is not only explained by the direct loading of items onto their common factors, but also by the indirect influence of factors correlated with their common factors. Factors are shifted in conceptual space (i.e., greater inter-factor correlations denoted by increasingly acute factor rotations) so as to replace some of the missing effect of any non-negligible loadings that simple-structure constrains to be zero. However, any missing effects not offset by increased inter-factor correlations, will serve to inflate correlations of unit-weighted subscale scores.

Concern about conceptual shifting due to imposing simple structure has been discussed in the measurement equivalency literature. Meredith and Horn (2001) recommended that factorial invariance studies be conducted on near-saturated structures prior to the imposition of markedly simpler structure. In near-saturated CFA models, a minimal set of marker items are constrained
to have zero loadings in order to define factors. All other items are free to load on each of the
defined factors. Because of the inclusion of all of the non-salient loadings and cross-loadings, the
fit of saturated CFA models should approximate the fit of corresponding EFA models. The
constraints are chosen by examining the EFA solution to identify items that have salient loadings
on one factor and near-zero (hyperplane) loadings on the other factors. By constraining the
hyperplane loadings to zero while freely estimating the salient loading, the CFA solution will
reproduce the EFA solution provided sufficient markers are used to define what does not load on
each factor\(^1\). Meredith and Horn’s concern was that the interpretation of any structure found to
be invariant should be qualitatively the same as the interpretation given the structure originally
obtained from EFA.

Similarly, single-factor models can yield different structure (i.e., interpretation) and fit when
compared to complex multidimensional models (multiple common-factors) because of the issue
of cross-loading. In a review of the measurement equivalency literature, Vandenberg (2002)
reminds us of the dictum that factors are defined in context, developing an example where
structure and fit are highly affected by a one-dimensional versus a multidimensional factor
model.

Shifting of the factor’s centroid in conceptual space is a qualitative idea in the sense that the
shift cannot be given any meaningful quantification, and the practical consequences depend on
the breadth (i.e., the specificity versus generality of constructs and their indicators) of the factor
and the scoring method. (For further discussion of general versus narrow factors, see Bagozzi &
Edwards, 1998) A broad factor with semantically dissimilar items (less specificity) may be more

\(^1\) Alternatively, factors can be defined in saturated models simultaneously with model identification by setting one or
more marker variables to unity for each factor. We prefer the less intuitive approach of defining a factor by what it is
not because this permits all salient loadings to be freely estimated and also places a restrictive simple-structure
model and a near-saturated model on a continuum where they differ only by the cutoff criterion (e.g., 0.30 versus
±0.05).
subject to re-interpretation if dissimilar items undergo large changes in the rank order of their loadings, while a narrow factor may be given the same interpretation even if all of its item loadings are substantially reordered. Note, changes in rank order of the loadings have little import for unit-weighted scores as long as each item continues to have a salient factor loading. However for a structural equations analysis where relations between the factor and other variables are estimated, a conceptual shift may not be ignorable.

The DBC may be susceptible to these types of structural issues when simple structure is imposed. Inspection of the DBC’s published loadings reveals many items with modest or low loadings. The two largest subscales (Disruptive, Self-Absorbed) have 27 and 31 items respectively. Six (22%) of the Disruptive subscale items have loadings of 0.45 or less, and 13 (42%) of the Self-Absorbed items have loadings of 0.46 or less, and at least 50% of the items on the remaining subscales have loadings of 0.46 or less. This structure is unstable across different cutoff criteria, in that retaining (or excluding) an item near the cutoff value may depend on chance. Also, the large number of items in the published structure that have factor loadings just above the cutoff suggests that each factor also had many item loadings just below the cutoff in the full EFA solution. With many modest loadings and many (unpublished) minor cross-loadings most of these items would be factorially complex.

Scoring

What is a significant difference between two scores on the same subscale? This issue is arguably even more salient for within-individual quantitative comparisons. What is a significant change? Nesselroade (2001) has argued that longitudinal studies attempting to assess the degree to which a person differs from themselves from one occasion to the next require greater measurement precision because within-person differences are likely to be smaller than between-
person differences. CFA models utilizing latent factors instead of manifest summary scores may provide such precision. It has been held for some time that unit-weighted and factor-weighted scores generally agree (Wackwitz & Horn, 1971), but these practices are based on continuous items and may be more problematic for polytomous items. Also, longitudinal contexts require consideration of the agreement in manifest and factor change scores rather than agreement of manifest and factor scores at any one occasion. Cattell (1966) called attention to the fact that a non-linear relationship between any observed scale and the underlying true score may significantly alter the correlation of changes in manifest scores with changes in true scale scores (i.e., the factor score). Appendix I provides an illustration based Cattell’s example (p. 366).

In screening applications with an emphasis on high negative predictive ability, it makes sense to retain items with modest loadings in the factor’s simple structure, but as noted above, it is unclear what impact this level of noise has on the context of quantitative comparisons. Observations falling into the false positive and false negative quadrants in Figure 2 illustrate this noise. The elongated cloud attests to a reasonably high correlation despite numerous misclassifications. It may also be more problematic when the item endorsement is infrequent and item distributions are highly skewed and dominated by zero (never), as is the case with the DBC. Finally, studies of agreement have generally looked at correlation and predictive ability – not level. High correlations are still obtained when there are modest discrepancies in the rank order of two variables, but correlation is relatively insensitive to the respective distributions of the two variables.
The Present Study

The issues raised here have particular relevance for the longitudinal Australian Child to Adolescent Development study (ACAD), which is in the unique position of having four longitudinal assessments (spanning 14 years) of the DBC in a population based sample of 570 children and youths with ID (hereafter denoted as ACAD epidemiological sample or ACAD-ES). Wave one of the ACAD-ES was part of the combined Dutch-Australian sample (hereafter denoted as DEKKER sample) used to develop the revised DBC structure and provide population norms (Dekker, Nunn et al., 2002). The genesis of the DBC had a diagnostic focus, but the ACAD study data provide an exceptional opportunity to investigate the natural history of BED as well as important associations with individual, family, and community covariates. This work is ongoing (Tonge & Einfeld, 2003), and interpretation of patterns of stability and change emerging from these studies can be facilitated by an examination of the DBC in regards to the above measurement and scoring issues.

Figure 2. Comparison unit-weighted and factor-weighted scoring methods using a cutoff value.
Dekker and colleagues’ (Dekker, Nunn et al., 2002) five-factor PCA solution accounted for only 43% of the variance, and the published DBC subscales were constructed by using cutoff criterion of 0.30 for the largest loading (or 0.40 for a secondary cross-loading). Most of the items (86 of 90) were retained with only four cross-loadings permitted in the whole structure. It is desirable to investigate how much noise this amount of unrelated variance (i.e., 57%) and potential factorial complexity introduces into quantitative comparisons of unit-weighted scores. Consequently, it is important to investigate the extent to which the DBC’s published structure is vulnerable to 1) inflated correlations among unit-weighted subscale scores, 2) problematic shifting of constructs in conceptual space when simple-structure CFA models are used, and 3) the degree of discrepancy between unit-weighted and factor-estimated scores.

Our goal is to identify the conditions under which aspects of the DBC’s subscales are vulnerable and invulnerable to the issues raised above. To examine the consequences of simple structure on model fit and factor interpretations, we will estimate CFA models for both the published simple structure as well as a more saturated model retaining many secondary loadings. To explore the consequences of manifest (i.e., unit-weighted) scoring versus factor-estimated scoring we will examine the sensitivity and specificity of manifest scores under the assumption that the factor-estimated scores in our near-saturated model represent true scores. Finally, we will inspect the inter-factor correlations for both CFA models and compare them to the correlations among unit-weighted subscale scores.

Several expectations follow from inspection of the DBC’s published structure. The modest and small magnitude of many of the published loadings suggests that many items may have non-

---

2 It is interesting to note that the developers of a similar instrument, the Reiss Screen for Maladaptive Behavior (RSMB) (Reiss & Valenti Hein, 1994), chose to construct ten subscales (i.e., extract 10 factors) but to retain only the five highest loading items for each subscale. Although the RSMB was intended for screening purposes, if the loadings of the five retained items were generally high and homogeneous, more precise quantitative comparisons would be facilitated – possibly at the cost of lower screening sensitivity.
negligible minor loadings and cross-loadings. For this reason the fit of the CFA simple-structure model is expected to be poor, and imposing simple-structure is expected to account for most of the misfit. Consequently, the fit of the near-saturated CFA model is expected to be good, and congruence with the published structure should be high. Some shift of the factors in conceptual space is expected in the simple-structure model as increases in the inter-factor correlations are expected to offset the removal of minor cross-loadings. It is expected that correlations of manifest subscale scores will be inflated, but it is not clear how much inflation or how uniform any inflation will be. Finally, there is no expectation about the performance of scoring procedures. The potential for discrepancy is easily articulated, but it is unclear whether these biases will aggregate or cancel out. These analyses must be considered exploratory.

Methods

Sample

The Australian Child to Adult Development study (ACAD) epidemiological cohort (n=570) was recruited in 1991 from all health, education and family agencies that provide services to children with mental retardation of all levels whose families lived in the census districts of Murray-Murrumbidgee, Grafton, Ryde-Hunters Hill, Sutherland and Illawarra in New South Wales, Australia, and the Dandenong-Westernport region in Victoria, Australia. The sampling procedure is described in more detail elsewhere (Einfeld & Tonge, 1992; Tonge & Einfeld, 1991). This cohort has been followed longitudinally for 14 years. The ascertainment process provided a virtually complete elaboration of young people (4-22 years at the first wave) with moderate and more severe mental retardation, but as with other studies, some young people with mild mental retardation from these regions were not identified. The mean age in years of the
cohort at was 12.1 ($SD = 4.38$). This cohort is representative of the general Australian community in terms of mix of social class, ethnic diversity, and rural-urban environment.

The ACAD-ES, a subsample comprising 570 of the 1536 children in the DEKKER sample, has some differences in age and ID distributions when compared to the full DEKKER sample. The mean age in the DEKKER sample was 12.0 ($SD=4.0$). The DEKKER sample contained a smaller percentage of children younger than six years. Percentages of youths with mild, moderate, severe, and profound ID are 32.0%, 41.0%, 22.7%, and 4.3% in this subsample versus 43.6%, 30.6%, 16.0%, 3.1% in the full sample; so there is a greater percentage of children with moderate and severe ID in our subsample.

**Measures**

*Developmental Behavior Checklist (DBC)*: The DBC (S.L. Einfeld & B.J. Tonge, 1995, 2002) is the key measure of psychopathology for young people with ID aged 4-18 years. It is a 96 item instrument which is completed by parents or other primary caregivers (Primary Care Version: DBC-P). A Total Behavior Problems Score (TBPS) can be calculated, along with scores on five factor-analytically derived subscales. The subscales allow for a description of five dimensions of disturbance: Disruptive/Antisocial, Anxiety Problems, Communications Disturbance, Self Absorbed Behavior, and Social Relating Problems.

**Missing Data**

A two fold approach to missing data was taken. First, based on follow-up contact with respondents and judgments by clinical staff, isolated and short sequences of missing item responses were imputed as zero. A series of follow-up contacts ascertained that respondents often failed to complete items when the answer was zero (never). This imputation was not utilized for sequences of 4 or more missing item responses. All remaining missing data were
handled by the MPlus maximum likelihood model-based (missing at random, MAR) capability. This is an advance over previous analyses of these data in which zero-imputation was used for all missing item responses. In particular, 154 subjects missing data on item #78 (Stands too close) were not imputed to zero because this missingness was due to the use of an early DBC-P version in which this item was not included (Overaffectionate and Stands too Close were initially one item).

*Polychoric correlations and MPlus thresholds*

All models are estimated with MPlus 3.12 (Muthen & Muthen, 1998-2005) which offers broad support for analyses of polytomous items such as the DBC’s response options (i.e., never, sometimes, always) including estimation with missing data, and robust estimators. Under assumptions of continuous and normally distributed indicators (i.e., Pearson product-moment correlations), polytomous indicators have poor distributional properties that may introduce serious misspecification problems or may bias fit statistics. The distributions for these indicators, even when symmetrical, can only offer a poor approximation of a normal distribution, and as with the ACAD-ES, are often very asymmetric (e.g., zero dominated). Prior PCA study of the DBC (Dekker, Koot, van der Ende, & Verhulst, 2002) used NOVAX, a stand-alone package by Niels Waller that implements polychoric (ordinal-ordinal) correlations.

The MPlus implementation combines polychoric correlations with the weighted least squares estimator (WLS), a variant of Brown’s (1984) asymptotic distribution-free estimator that performs well for smaller sample sizes. Polychoric correlations are actually performed using estimated thresholds instead of observed response frequencies. Unobserved continuous response variables are conceptualized to underlie each observed polytomous response. Cutpoints on this distribution are estimated such that the observed response frequencies would result. The MPlus
implementation provides for simultaneous estimation of these cutpoints their subsequent
polychoric correlation, and the factor-analytic model. (For further discussion and references see
Bontempo & Hofer, in press). The MPlus implementation additionally provides several robust
variants of the WLS estimator. Pairwise methods used to produce the polychoric correlation
matrix can result in a non-positive-definite information matrix which cannot be inverted in the
course of WLS estimation. This is most likely to happen when the sample size is small and the
indicators are highly skewed.\(^3\) The WLSM and WLSMV estimators (Muthén, du Toit, & Spisic,
1997) do not require the information matrix to be inverted. In simulation studies the WLSMV \(\chi^2\)
fit statistic has outperformed the WLS \(\chi^2\) fit statistic.\(^4\) All models in our study are estimated with
the WLSMV estimator because the information matrix was not positive-definite.

Factors versus principal components

Given the relatively low percent of variance explained in the DEKKER sample (43%), we
extract factors instead of principal components. The current five-factor structure of the DBC is
based on principal component extraction. The variant of the WLS estimator we employ
constitutes principal axis extraction. Methodology will be a potential explanation for any
observed lack of congruence between our models and the published structure.

\(^3\)Personal Communication from Linda K Muthén, Jan 16, 2004.
(http://www.statmodel.com/discussion/messages/23/282.html?#POST2267)
Results

Structure

To assess misfit and conceptual shift due to the imposition of simple structure, both simple-structure CFA models based on the published structure, as well as a near-saturated CFA model were estimated for the ACAD-ES. Mean and variance structures were estimated. Both CFA models were identified by setting each of the five common factors to have unit variance and mean zero. Taken together, the fit of these models and any marked departure from the published pattern of loadings are informative about the robustness of the published structure to sampling variation and statistical methodology, as well as the consequences of imposing simple structure. Ideally, the PCA solution from the validation study (DEKKER sample) would be used to select marker items for the near-saturated CFA model, and near-saturated and simple-structure CFA models of the DEKKER sample would be compared. However CFA software for polytomous items was not available at that time, and the full PCA solution for the DEKKER sample is not available at the present time. Marker items for our near-saturated model are thus based on a five-factor EFA solution for the ACAD-ES. Markers items for each factor were established by selecting only those items in the EFA solution with loadings within ±0.05 of zero.

The number of constrained marker items (i.e, a loading fixed to zero) for each factor in the near-saturated model varied, but each factor had considerably fewer constraints compared to the simple-structure model. Six (as opposed to the 59 under the published simple structure) marker items were used for self absorbed: 5, 22, 39, 56, 71, 95. Eleven (as opposed to 63) marker items were used for disruptive/antisocial: 3, 10, 23, 28, 45, 55, 57, 61, 67, 71, 89. For communication disturbance, 14 (as opposed to 77) marker items were used: 4, 9, 24, 26, 27, 31, 40, 46, 48, 61, 67, 68, 70, 86. For social relations, 18 (as opposed to 80) marker items were used: 6, 8, 9, 10, 13,
27, 30, 31, 33, 40, 55, 62, 63, 72, 76, 77, 82, 87. For anxiety, 23 (as opposed to 81) marker items were used: 2, 3, 4, 5, 8, 10, 18, 25, 27, 32, 27, 44, 50, 54, 55, 56, 59, 61, 76, 78, 82, 87, 94.

Fit of the simple-structure CFA model was poor ($\chi^2 (272)^5 = 1076.0, p < .001; \text{CLI} = 0.78; \text{TLI} = 0.85; \text{RMSEA} = 0.072; 272 \text{ free parameters}$), but the fit of the near-saturated CFA model was good ($\chi^2 (294)^5 = 534.6, p < .001; \text{CLI} = 0.93; \text{TLI} = 0.96; \text{RMSEA} = 0.039; 565 \text{ free parameters}$). The simple structure model explained 32.8% of the variance. The near-saturated model explained 42.2% of the variance. The EFA model ($\chi^2 (295) = 578.0, p < .001; \text{RMSEA} = 0.041$) explained 44.3% of the variance, which compares favorable with the 43.7% of the variance originally reported (Dekker, Nunn et al., 2002). The similar RMSEA value and percent of variance explained for our EFA and near-saturated CFA models is evidence that the poor fit of the simple-structure CFA model is due to excess constraints and not due to sampling variation or statistical methodology. Approximately 21% of the variance explained by the EFA model is unexplained by the simple-structure CFA model.

Substantive comparability (congruence) between our two CFA solutions and the published structure depends on the degree of discrepancy in magnitude and rank order of the larger loadings (i.e., those defining each factor), any alterations in the pattern of item-factor associations, and differences in the inter-factor correlations. However, no statistic is available to quantify congruence and a judgment call must be made. Table 1a and 1b show the correspondence of the published loadings (Dekker, Nunn et al., 2002; Einfeld & Tonge, 2002) with our two CFA models. Discrepancies for the near-saturated model may reflect sampling variation or methodology. Discrepancies in the simple-structure model, especially when not reflected in the near-saturated model, can reasonably be seen as a shift of the factor’s centroid in

---

5 Adjusted degrees of freedom and $\chi^2$ for the WLSMV estimator are estimated according to a formula given in the Mplus Technical Appendices at www.statmodel.com.
conceptual space. Our EFA model is not tabulated here, but was very similar to our near-saturated CFA model except that slightly more item-factor associations were retained (i.e., exceeded the cutoff of 0.30). Table 3 shows the change in inter-factor correlations between the near-saturated and simple structure model CFA models, as well as the correlations among unit-weighted subscale scores.

Table 1a and Table 1b show DBC-P item short names, published PCA loadings, and loadings for the simple-structure and near-saturated CFA models. Item short names and DBC-P item number are shown in the 1st column. The 2nd column shows the published loading from the DBC manual. The 3rd column shows the loadings obtained in our simple-structure CFA, and the 4th column shows the near-saturated CFA loadings. The tabulation used here is thus an expanded item-factor view because it shows each item in the published structure (even if it did not meet the 0.30 cutoff criteria in our models), secondary cross-loadings in the 0.30 to 0.40 range which were excluded by Dekker, as well as any new salient (>0.30) item-factor associations in our seen in our near-saturated CFA (and EFA) model. To highlight potential conceptual shift denoted by changes in loading magnitude and/or rank order, three types of discrepancies are indicated by bold text and superscript notation. Items where the near-saturated CFA loading differs from the published loading by more than ± 0.1 are superscripted with “a”. Items where the simple-structure CFA loading differs from the published loading by more than ± 0.2 are
Table 1a. Simple-structure and near-saturated five-factor CFA solutions for the ACAD epidemiological sample.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Item Loadings</th>
<th>Item Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td># items (# retained)</td>
<td>Item Short Name [DBC-P #]</td>
<td></td>
</tr>
<tr>
<td>Disruptive</td>
<td>27 items (25 retained)</td>
<td></td>
</tr>
<tr>
<td>Abusive [#04]</td>
<td>.86</td>
<td>.66&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Lies [#83]</td>
<td>.81</td>
<td>.59&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Stubborn [#74]</td>
<td>.78</td>
<td>.75</td>
</tr>
<tr>
<td>Manipulates [#87]</td>
<td>.73</td>
<td>.73</td>
</tr>
<tr>
<td>Steals [#73]</td>
<td>.73</td>
<td>.65</td>
</tr>
<tr>
<td>Bossy [#93]</td>
<td>.71</td>
<td>.60</td>
</tr>
<tr>
<td>Impulsive [#37]</td>
<td>.68</td>
<td>.70</td>
</tr>
<tr>
<td>Kicks [#40]</td>
<td>.67</td>
<td>.71</td>
</tr>
<tr>
<td>Tantrums [#31]</td>
<td>.64</td>
<td>.71</td>
</tr>
<tr>
<td>Impatient [#35]</td>
<td>.64</td>
<td>.77</td>
</tr>
<tr>
<td>Irritable [#38]</td>
<td>.62</td>
<td>.78</td>
</tr>
<tr>
<td>Jealous [#39]</td>
<td>.61</td>
<td>.58</td>
</tr>
<tr>
<td>Whines [#95]</td>
<td>.59</td>
<td>.67</td>
</tr>
<tr>
<td>Hides [#32]</td>
<td>.57</td>
<td>.57</td>
</tr>
<tr>
<td>Attention seek [#53]</td>
<td>.56</td>
<td>.70</td>
</tr>
<tr>
<td>Fires [#43]</td>
<td>.56</td>
<td>.53</td>
</tr>
<tr>
<td>Not capable [#77]</td>
<td>.54</td>
<td>.50</td>
</tr>
<tr>
<td>Easily led [#20]</td>
<td>.53</td>
<td>.47</td>
</tr>
<tr>
<td>Talks fast [#81]</td>
<td>.53</td>
<td>.47</td>
</tr>
<tr>
<td>Throws [#86]</td>
<td>.51</td>
<td>.40</td>
</tr>
<tr>
<td>Mood [#47]</td>
<td>.50</td>
<td>.79&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Refuses to go [#59]</td>
<td>.45</td>
<td>.44</td>
</tr>
<tr>
<td>Noisy [#49]</td>
<td>.45</td>
<td>.60</td>
</tr>
<tr>
<td>Runs away [#14]</td>
<td>.42</td>
<td>.31</td>
</tr>
<tr>
<td>Overactive [#50]</td>
<td>.41</td>
<td>.37</td>
</tr>
<tr>
<td>Tense [#85]</td>
<td>.39</td>
<td>.58</td>
</tr>
<tr>
<td>Esteem [#41]</td>
<td>.37</td>
<td>.44</td>
</tr>
<tr>
<td>Unhappy [#01]</td>
<td>.43</td>
<td></td>
</tr>
<tr>
<td>Screams [#66]</td>
<td>.42</td>
<td></td>
</tr>
<tr>
<td>Cries [#11]</td>
<td>(A .42)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item Loadings</td>
<td>Item Loadings</td>
</tr>
<tr>
<td># items (# retained)</td>
<td>Item Short Name [DBC-P #]</td>
<td></td>
</tr>
<tr>
<td>Self-Absorbed</td>
<td>31 items (27 retained)</td>
<td></td>
</tr>
<tr>
<td>Non food [#21]</td>
<td>.85</td>
<td>.63&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Hums [#34]</td>
<td>.78</td>
<td>.65</td>
</tr>
<tr>
<td>Mouths [#10]</td>
<td>.75</td>
<td>.54&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Bites [#08]</td>
<td>.67</td>
<td>.53</td>
</tr>
<tr>
<td>Soils [#70]</td>
<td>.66</td>
<td>.55</td>
</tr>
<tr>
<td>Danger [#55]</td>
<td>.65</td>
<td>.59</td>
</tr>
<tr>
<td>Hits self [#33]</td>
<td>.65</td>
<td>.58</td>
</tr>
<tr>
<td>String [#44]</td>
<td>.63</td>
<td>.63</td>
</tr>
<tr>
<td>Smells [#64]</td>
<td>.62</td>
<td>.63</td>
</tr>
<tr>
<td>Lights [#72]</td>
<td>.60</td>
<td>.60</td>
</tr>
<tr>
<td>Movements [#60]</td>
<td>.59</td>
<td>.67</td>
</tr>
<tr>
<td>Stares [#68]</td>
<td>.59</td>
<td>.54</td>
</tr>
<tr>
<td>Wanders [#94]</td>
<td>.56</td>
<td>.67</td>
</tr>
<tr>
<td>Flicks [#25]</td>
<td>.55</td>
<td>.62</td>
</tr>
<tr>
<td>Throws [#86]</td>
<td>.54</td>
<td>.41</td>
</tr>
<tr>
<td>Bangs head [#06]</td>
<td>.54</td>
<td>.43</td>
</tr>
<tr>
<td>Urinates [#92]</td>
<td>.52</td>
<td>.49</td>
</tr>
<tr>
<td>Gorges [#27]</td>
<td>.52</td>
<td>.56</td>
</tr>
<tr>
<td>Overactive [#50]</td>
<td>.46</td>
<td>.42</td>
</tr>
<tr>
<td>Runs away [#14]</td>
<td>.46</td>
<td>.36</td>
</tr>
<tr>
<td>Laughs [#42]</td>
<td>.46</td>
<td>.64</td>
</tr>
<tr>
<td>In public [#46]</td>
<td>.46</td>
<td>.51</td>
</tr>
<tr>
<td>Clothes [#76]</td>
<td>.44</td>
<td>.60</td>
</tr>
<tr>
<td>Aloof [#03]</td>
<td>.43</td>
<td>.28</td>
</tr>
<tr>
<td>Grinds [#29]</td>
<td>.43</td>
<td>.24</td>
</tr>
<tr>
<td>Pain [#88]</td>
<td>.42</td>
<td>.51</td>
</tr>
<tr>
<td>Screams [#66]</td>
<td>.41</td>
<td>.72&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Attention [#09]</td>
<td>.39</td>
<td>.61&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Posture [#30]</td>
<td>.36</td>
<td>.44</td>
</tr>
<tr>
<td>Twitches [#24]</td>
<td>.33</td>
<td>.44</td>
</tr>
<tr>
<td>Noisy [#49]</td>
<td>(D .45)</td>
<td></td>
</tr>
<tr>
<td>Feelings [#18]</td>
<td>(SR .38)</td>
<td></td>
</tr>
<tr>
<td>Distracted [#19]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Near-saturated loading differs from published loading by more than 0.1
<sup>b</sup> Simple-structure loading differs from published loading by more than 0.2
<sup>c</sup> Dropped item. Near-saturated loading is less than .30, item may or may not load on another factor.
Table 1b. Simple-structure and near-saturated five-factor CFA solutions for the ACAD epidemiological sample.

<table>
<thead>
<tr>
<th>Scale</th>
<th>Item Loadings</th>
<th>Item Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td># items (# retained)</td>
<td>Item Short Name [DBC-P #]</td>
<td># items (# retained)</td>
</tr>
</tbody>
</table>

### Anxiety
9 items (5 retained)

<table>
<thead>
<tr>
<th>Item Short Name [DBC-P #]</th>
<th>Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alone [#16]</td>
<td>.60</td>
</tr>
<tr>
<td>Familiar [#22]</td>
<td>.60</td>
</tr>
<tr>
<td>Nightmares [#30]</td>
<td>.52</td>
</tr>
<tr>
<td>Fears [#23]</td>
<td>.49</td>
</tr>
<tr>
<td>Cries [#11]</td>
<td>.42</td>
</tr>
<tr>
<td>Shy [#75]</td>
<td>.37</td>
</tr>
<tr>
<td>Appetite [#45]</td>
<td>.35</td>
</tr>
<tr>
<td>Changes [#91]</td>
<td>.32</td>
</tr>
<tr>
<td>Food fads [#26]</td>
<td>.30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Item Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tense [#85] (D .39)</td>
</tr>
<tr>
<td>Vomits [#52] (SR .38)</td>
</tr>
<tr>
<td>Grinds [#29] (SA .43)</td>
</tr>
<tr>
<td>Screams [#66] (SA .41)</td>
</tr>
</tbody>
</table>

**Social Relations**
10 items (8 retained)

<table>
<thead>
<tr>
<th>Item Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underactive [#48] (D .59)</td>
</tr>
<tr>
<td>Affection [#17]</td>
</tr>
<tr>
<td>Unhappy [#01]</td>
</tr>
<tr>
<td>Sleeps much [#69]</td>
</tr>
<tr>
<td>Cuddled [#61]</td>
</tr>
<tr>
<td>Aloof [#03]</td>
</tr>
<tr>
<td>Eye contact [#02]</td>
</tr>
<tr>
<td>Feelings [#18]</td>
</tr>
<tr>
<td>Vomits [#52]</td>
</tr>
<tr>
<td>Her own [#57]</td>
</tr>
</tbody>
</table>

**Dropped item. Near-saturated loading is less than .30, item may or may not load on another factor.**

### Communication Disturbance
13 items (13 retained)

<table>
<thead>
<tr>
<th>Item Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routine [#05]</td>
</tr>
<tr>
<td>Echo [#62]</td>
</tr>
<tr>
<td>Pronouns [#13]</td>
</tr>
<tr>
<td>Talks to self [#82]</td>
</tr>
<tr>
<td>Repeats [#63]</td>
</tr>
<tr>
<td>Whispers [#71]</td>
</tr>
<tr>
<td>Elated [#89]</td>
</tr>
<tr>
<td>Obsessed [#28]</td>
</tr>
<tr>
<td>Mechanical [#54]</td>
</tr>
<tr>
<td>Doesn't mix [#56]</td>
</tr>
<tr>
<td>Preoccupied [#58]</td>
</tr>
<tr>
<td>Overaffectionate [#51]</td>
</tr>
</tbody>
</table>

**Near-saturated loading differs from published loading by more than 0.1**

**Simple-structure loading differs from published loading by more than 0.2**

**Dropped item. Near-saturated loading is less than .30, item may or may not load on another factor.**
superscripted with “b”. These criteria are arbitrary and are for illustrative purposes. Items disassociating with a factor by not meeting the 0.30 cutoff criteria in our near-saturated CFA model are superscripted with “c”. Finally, items newly associating with an item by loading in excess of 0.30 in our near-saturates CFA model are shown below the published items.

There is strong evidence that a congruent rotated factor structure obtains for the Self-Absorbed and the Disruptive/Antisocial subscales, and weaker evidence for congruence of the Social-Relating, the Communication Disturbance, and the Anxiety subscales. All five factors continue to be defined by several high loadings for the same items that define the published subscales. Consistent with expectations, the most disparity is seen in the simple-structure CFA solution. For the Disruptive, Self-Absorbed, and Communication Disturbance subscales there is little or no change in rank order of the loadings, however Communication Disturbance exhibits new associations with a number of additional items. The Social Relations and the Anxiety subscales show more changes in magnitude and rank order of loadings. As expected, the large majority of changes in item-factor association (for any subscale) occur for items with low to moderate (i.e., .30 to .45 range) factor loadings.

The published loadings for the Self-Absorbed subscale items are very similar to the structure in our CFA models. In the near-saturated CFA model 27 of the 31 items are retained. Four items, Grinds (#29), Pain (#88), Posture (#90), and Twitches (#24) have loadings slightly below the 0.30 cutoff. Rank order of loadings is also very consistent with the published loadings. Only Stares (#68) is more than 0.10 lower than the corresponding published loading and Overactive (#50) is more than 0.10 higher. There is only minor shuffling of a few additional loadings in the simple-structure CFA model. Non food (#21) and Mouths (#10) load more than 0.20 lower, while Excited (#07), Screams (#66), and Attention (#09) load more than 0.20 higher.
Screams and Attention reflect the greatest shift in that they reposition from among the very lowest loading items to the highest or near highest loading items. Three new items (Noisy, #45; Feelings, #18; Distracted #19) loaded above 0.30 in our EFA and near-saturated CFA models.

Published loadings for items on the Disruptive/Antisocial subscale also correspond very well with the structure in our CFA models. In our near-saturated CFA model 26 of the 26 items are retained. Two items, Overactive (#50) and Esteem (#41) have loadings just below the 0.30 cutoff, but in total 13 items have loadings more than 0.10 lower than the corresponding published loading, although changes in rank order are generally minor. Only Mood (#47) has a loading more than 0.10 higher. Also, several of these items would not be included on the Disruptive subscale under the original rules because they are now cross-loadings in the 0.30 to 0.40 range. Only two items, Impulsive (#37) and Mood (#64) move more than a few positions in rank order. In the simple-structure CFA model, Abusive (#04) and Lies (#83) have loadings more than 0.20 lower than the published loading, and Mood (#47) has a higher. The only large discrepancy is with Mood, which shifts from a moderate loading (.50) to the highest loading (.79). Also, three previously unassociated items (Unhappy, #1; Screams, #66; Cries, #11) loaded above 0.30 in our EFA and near-saturated CFA models.

The three smaller factors corresponding to the Communication Disturbance, Social relations, and Anxiety subscales (Table 1b) show varying degrees of congruence, but have proportionally more discrepancies than the two larger factors. Discrepancies are seen in both the near-saturated and simple-structure models. For the factors corresponding to the Communication Disturbance and the Social Relations subscales, the majority of the items are retained in the EFA and near-saturated CFA model, but there are still substantial changes in the rank order of
loadings and/or many newly associating items. For Anxiety, only five of nine items are retained, but there is less reordering of items.

The factor corresponding to the Communication Disturbance subscale is the only one of the five factors that continues to associate with all of its corresponding published subscale’s items. However, this factor may also evidence the largest conceptual shift seen in our models as ten items newly associate with this factor (i.e., now have loadings of 0.30 or greater). Only two of these new items (Hides, #32; Her Own, #57) would have failed to meet the 0.40 cross-loading criterion. The remaining eight would have been retained as a loading (> .30) or cross-loading (> .40) under the criteria used by Dekker (Dekker, Nunn et al., 2002), and most of these new items do not invoke the idea of communication (e.g., Easily Led, #20; Not Capable, #77; Lights, #72). The majority of these new associations are from the Disruptive subscale. In the near-saturated CFA model, the rank-order of the six largest published loadings reproduces very well. These items are now rank 2nd through 7th in the same order (excluding the newly associated items).

The exception is Stands (#78), which has the lowest published loading (.31), yet has the highest loading in both the near-saturated (.62) and simple-structure (.63) models. This is most likely the result of inappropriate zero-imputation used in prior work (i.e., the 154 cases where the item was not present on the early versions of the DBC-P). It seems likely that the ascendancy of Stands is linked to all the new item associations. In the simple-structure model the conceptual shift is also seen. The highest three published item loadings (Routine, #05; Echo, #05; Pronouns, #62) are now ranked 13th, 6th, and 11th; and three of the lowest four loadings (Preoccupied, #58; Stands, #78; Doesn’t Mix, #56) now rank in the top five. Non-verbal items now dominate the factor.
The factor corresponding to the Social Relations subscale has eight of ten items from the published subscale continuing to associate with the factor. Most items loadings change rank order, but the practical consequence may be negligible because the items are similar. Underactive (#48), the highest published loading (.59), falls to 10th place (.25) in the rank order of the simple-structure model, while the item with the lowest published loading (Her own, #57), has highest loading in the simple-structure model. Two items (Sleeps much, #69; Vomits, #52) fall below the 0.30 cutoff in the near-saturated model, although Sleeps Much is just below the cutoff. Six previously unassociated items loaded above 0.30, but only three (Preoccupied, #58; Fires, #43; Twitches, #34) would have been retained as a loading (> .30) or cross-loading (> .40).

Finally, the factor corresponding to the Anxiety subscale has only five of the nine published items continuing to associate with the factor, but a similar structure is found. The three highest published loadings (Alone, #16; Familiar, #22; Nightmares, #30) are reproduced with the same rank order in the near-saturated model, but not in the simple-structure model. Here too however, the practical consequences may be small because the construct is narrow. Additionally, the four items disassociating in the near-saturated model still load in the .22 to .27 range. There are also four previously unassociated items now loading in the .34 to .39 range in the near-saturated model. Three of these (Tense, #85; Vomits, #52; Grinds, #29) would have been retained as a loading and seem conceptually similar to the original items. In the simple-structure model, Changes (#91) moves from one of the lowest loadings (.32) to the highest loading item (.69), but Alone (#16), the highest published loading (.60), still has a high loading (.57).
Scoring.

To explore the implications of using manifest (unit-weighted) versus factor-estimated scoring, scores were computed with both methods, ranked into quintiles, and cross-tabulated to assess the degree of concordance. Factor scores were generated from our near-saturated CFA model, and manifest scores were based on the published subscales. Figure 1 illustrates discrepancies in Self-Absorbed and Anxiety based on selection at the 75th percentile. At each level of unit-weighted score (vertical axis), there is considerable range of factor-weighted scores (horizontal axis). The range of factor scores is much wider for Anxiety than for Self-Absorbed. Table 2 shows the percentage of correct and incorrect classification within each quintile for each subscale’s range of scores. Also tabulated is the total percentage of cases that are seriously misclassified (defined as more than one quintile of discrepancy), and the Goodman and Kruskal Gamma statistic (Agresti, 1996), which is a measure of association that is sensitive to rank-order changes.

The Disruptive and Self-Absorbed subscales show only modest discrepancies due to scoring method. The total percent of cases seriously discrepant is 1.9% and 1.6% respectively. However there is considerable minor discrepancy. For the Disruptive subscale (gamma = .79), the 2nd, 3rd, and 4th quintiles show only 56.1%, 47.4%, and 57.9% of concordant pairs. For the Self-Absorbed subscale (gamma = .82) the 2nd, 3rd, and 4th quintiles show only 56.1%, 50.4%, and 71.9% of concordant pairs. This subscale shows the smallest effect of scoring. Figure 2 contrasts scoring discrepancy for the Self-Absorbed subscale with the Anxiety subscale, which showed the most discrepancies.

---

6 Illustrations would vary with quartiles or deciles versus quintiles. Increased discrepancy is not necessarily monotonic with more bins, but also depends on the two distributions and the number of cases near bin boundaries. However, the intent here is exploratory.

7 The near-saturated CFA is the closest approximation to an EFA solution for which no software is available to provide factor scores.
The Communication Disturbance and Social Relations subscales show a comparably larger effect of scoring method. Shorter scales are more sensitive because each item holds proportionately more weight. The total percent of cases seriously discrepant is 4.1% and 5.6% respectively. There is considerable minor discrepancy. For the Communication Disturbance subscale ($\gamma = .75$), the 2$^{nd}$, 3$^{rd}$, and 4$^{th}$ quintiles show only 44.7%, 55.3%, and 45.5% of concordant pairs. For the Social Relations subscale ($\gamma = .76$), the 2$^{nd}$, 3$^{rd}$, and 4$^{th}$ quintiles show only 38.3%, 46.5%, and 54.1% of concordant pairs.

The Anxiety subscale ($\gamma = .61$) shows the most serious discordance. The total percent of cases seriously discrepant is 16.0%. In some cases, outliers are more than two quintiles discrepant. The 2$^{nd}$, 3$^{rd}$, and 4$^{th}$ quintiles show only 19.3%, 33.9%, and 34.2% of concordant pairs.

Subscale and Inter-Factor Correlations.

To provide further insight into the scoring discrepancies noted above, as well as any structural changes seen in the simple-structure CFA model, it is important to also consider changes in subscale and inter-factor correlations. Table 3 shows the inter-factor correlations for the near-saturated CFA and simple-structure CFA models, as well as for

![Figure 3. Self-Absorbed and Anxiety Scoring Discrepancies](image-url)
unit-weighted subscale scores. Under oblique (e.g., promax) rotation, item variance is explained not only by its associated common-factor, but also by the other common-factors that correlate with the items’ common factor. The inter-factor correlations in our near-saturated CFA model are small to modest (.02 to .39). This is consistent with the .12 to .34 range reported for the DEKKER sample.

The inter-factor correlations for both the simple-structure CFA model and the unit-weighted scores are two to three times larger and show distortion. In the CFA model inter-factor correlations increase to offset the numerous cross-loadings fixed to zero. In a similar manner, the larger correlations among the unit-weighted scores reflect the factorial complexity of many of the DBC’s items. The correlation between Disruptive/Antisocial and Social Relations shows the most distortion, increasing from near zero to 0.34. The correlation between Self-Absorbed and Social Relations also shows a disproportionate increase.
Table 2. Classification and mis-classification by quintiles of unit-weighted and factor-weighted scores.

<table>
<thead>
<tr>
<th>Factor-weighted Quintile</th>
<th>% Serious&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Success</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Gamma&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Disruptive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>76.1</td>
<td>19.3</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>23.0</td>
<td>56.1</td>
<td>28.1</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.9</td>
<td>21.9</td>
<td>47.4</td>
<td>20.2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2.6</td>
<td>21.9</td>
<td>57.9</td>
<td>17.7</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.9</td>
<td>19.3</td>
<td>81.4</td>
<td></td>
</tr>
<tr>
<td>Self-Absorbed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>86.8</td>
<td>17.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>12.3</td>
<td>56.1</td>
<td>30.1</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.9</td>
<td>22.8</td>
<td>50.4</td>
<td>13.2</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3.5</td>
<td>18.6</td>
<td>71.9</td>
<td>15.0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.9</td>
<td>12.3</td>
<td>85.0</td>
<td></td>
</tr>
<tr>
<td>Communication-Disorder</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>77.0</td>
<td>17.5</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>20.4</td>
<td>44.7</td>
<td>22.8</td>
<td>2.7</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.9</td>
<td>31.6</td>
<td>55.3</td>
<td>31.3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>1.8</td>
<td>6.1</td>
<td>17.5</td>
<td>45.5</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.6</td>
<td>20.5</td>
<td>72.8</td>
<td></td>
</tr>
<tr>
<td>Social-Relations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>73.5</td>
<td>21.7</td>
<td>6.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>19.5</td>
<td>38.3</td>
<td>15.8</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>7.1</td>
<td>30.4</td>
<td>46.5</td>
<td>28.1</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>9.6</td>
<td>29.8</td>
<td>54.4</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>1.8</td>
<td>15.8</td>
<td>74.1</td>
<td></td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>66.7</td>
<td>35.1</td>
<td>12.2</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>17.5</td>
<td>19.3</td>
<td>19.1</td>
<td>10.5</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>12.3</td>
<td>28.9</td>
<td>33.9</td>
<td>30.7</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3.5</td>
<td>14.0</td>
<td>26.1</td>
<td>34.2</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2.6</td>
<td>8.7</td>
<td>21.1</td>
<td>67.9</td>
</tr>
</tbody>
</table>

<sup>a</sup> Mis-classification by more than one quintile.

<sup>b</sup> Also called Goodman and Kruskal's gamma, is a symmetric measure which varies from +1 to -1, based on the difference between concordant pairs. Gamma defines perfect association as weak monotonicity.
Discussion

The findings of this study can inform the use of the Developmental Behavior Checklist (DBC) (Einfeld & Tonge, 1992) in individual differences research, as well as facilitate investigations of the DBC’s measurement properties. The DBC’s intended purpose was for clinical screening. Decisions made during the construction and validation of a diagnostic measure are not necessarily the same decisions that would best serve an instrument designed for individual differences research. Retaining many items, including items with modest or small loadings and factorially complex items, may serve screening goals by increasing the instrument’s negative predictive ability. However, this may introduce problems for simple-structure confirmatory factor analyses (CFA) needed for some investigations of the instrument’s measurement properties, and may also add excess noise into subscale scores. Our examinations of the DBC’s structure and scoring procedures provide evidence that the two largest subscales (Disruptive/Antisocial, Self Absorbed) are performing well in CFA models and exhibit little scoring bias. Findings for the three smaller subscales (Communication Disturbance, Social Relations, Anxiety) highlight several issues that should be considered by researchers using these subscales for individual differences research.

Using the ACAD epidemiological sample to estimate two CFA models, one specifying the DBC’s published simple-structure and the other specifying a near-saturated structure, we were
able to investigate several structural and scoring issues. Congruence between our near-saturated CFA model and the published structure extracted with principal components analysis (PCA) provides some evidence that the DBC’s structure particularly two larger subscales, is robust to sampling variation and statistical methodology. The difference in fit between our near-saturated and simple-structure models provides further evidence that the poor fit of the simple-structure model is a decrement associated with imposing simple structure and not due to sampling or methodology. Any lack of congruence between our simple-structure model and our near-saturated model indicates how the factors are shifting in conceptual space, although the practical consequences of any shift depends on other factors such as the breadth of the latent construct and the particular items with discrepant loadings. Factor score estimates generated by our near-saturated CFA model permit an investigation of biases in unit-weighted subscale scores. Finally, comparison of the inter-factor correlations from our two CFA models with the correlations among unit-weighted subscale scores provides further insight into conceptual shifts due to imposing simple-structure or employing unit-weighted scoring.

In general, the findings confirmed our expectations about subscale structure congruency across models and about the precision of subscale scores. Because the ACAD epidemiological sample is a subsample (571 of 1536) of the DEKKER sample used to extract the DBC’s revised five-factor structure (Dekker, Nunn et al., 2002), we did not expect any substantial incongruence between our near-saturated CFA (and EFA) model and the published structure. As expected, our near-saturated CFA model (and comparable EFA model) fit well and evidenced good congruence when compared to the published subscale structure. The degree of incongruence did increase for the three smaller subscales (Communications Disturbance, Social Relations, Anxiety).
Also consistent with our expectations, the fit of the simple-structure CFA model was poor, the model accounted for substantially less variance, and inter-factor correlations were inflated. In addition, some re-ordering of item loadings (representing potentially problematic shifts in factor interpretation) was seen for the three smaller factors. We expected substantial misfit because the published structure contained many moderate and low loadings which suggested the likelihood that many constrained cross-loadings were not in the hyperplane. To recover some of the explained variance lost when minor cross-loadings are constrained to zero, estimation procedures can converge on qualitatively different factors with higher inter-factor correlations and a different rank order of item loadings. In our simple-structure CFA the inter-factor correlations were two to three times larger (e.g., .59 vs. .22). These findings are consistent with CFA studies of EFA derived personality structures where researchers have concluded that simple-structure is an unreasonable hypothesis for instruments with many items and broad latent constructs, or that more saturated models maybe needed, or that post hoc modifications (permitting some correlated residuals) should be used to offset the effect of constrained cross-loadings (Church & Burke, 1994; Raykov, 1998; Van Proooijen & Van Der Kloot, 2001).

Findings related to problems with unit-weighted scores were mixed. We were concerned that the high proportion of items with small loadings, as well as the relatively large amount (approximately 50%) of unexplained variance might distort unit-weighted subscale scores and inflate correlations among subscale scores. However, the Disruptive/Antisocial and Self Absorbed subscales showed high specificity and sensitivity and a strong concordance between factor scores and unit-weighted subscale scores. By contrast, increasing degrees of bias were observed for the three smaller subscales. For Anxiety, cross-tabulation showed that 16% of cases were discrepant in the rank order of unit-weighted and factor-weighted scores by more than two
quintiles. The correlations among unit-weighted subscale scores are inflated to about the same extent as the inter-factor correlations from the simple-structure model. In this case most of the inflation and distortion results from unrelated variance specific to each item instead of oblique rotation, but the factorial complexity of many items is the root cause in both cases.

Each of these findings may or may not impair the precision of individual difference research depending on the needs of the study and the methods employed. When individuals are compared to themselves over time high precision is generally needed. If mean levels across groups are being compared, lower precision may be adequate. If unit-weighted scores are being used conceptual shifting of factors is not a problem, although regression relationships between subscale scores and other variables will be distorted by the portion of the subscale score variance related to other subscales. If structural equation models containing latent factors are to be used, for example, the reversal of the highest and lowest loadings on the Social Relations factor [Underactive (#48), Her own (#57)] in our simple structure CFA model might be regarded as less important since the items both invoke non-social activity. By contrast, the reversal of the highest and lowest loadings on the Communications Disturbance factor [Routine (#5), Stands (#78)] might not be ignorable if the items are judged to have sufficiently distinct meanings in the context of disturbed communications. Some intervention, possibly omitting the item about standing to close, may be needed to preserve the interpretation of the communications disturbance factor in these structural models.

The number of discrepant items falling below and rising above the 0.30 cutoff in our near-saturated CFA model has implications for the precision of subscale scores. Given that the ACAD epidemiological sample is a subsample of the DEKKER sample with only small differences in age and ID distributions, the structure of problem behaviors and emotions in this model could be
expected to be highly concordant with the published structure. These discrepancies are plausibly
due to a combination of sampling variation, statistical methodology, and the improved missing
data treatment of item #78 (Stands). However these discrepancies suggest that items with low
loadings may have been included or excluded from each subscale by chance. Eliminating these
items would trade sensitivity for specificity.

Problematic conceptual-shifting of a factor’s interpretation in simple-structure CFA models
would be a concern for researchers investigating the measurement equivalency of the DBC’
subscales across age, ID, or cultural groups. It is of little value to find that some construct other
than the one assessed by the unit-weighted score is invariant across groups. In addition, these
procedures require a reasonably well fitting CFA model to use as a baseline for testing fit
decrements in more restricted models constraining measurement parameters to be equivalent
across groups (For review see, Bontempo & Hofer, in press; Vandenberg & Lance, 2000).

Our findings suggest that only the Communication Disturbance subscale exhibits an
unacceptable conceptual-shift. Our near-saturated model suggests that this shift is partly related
to the missing data issue for the item about standing too close (Stands, #78), but whatever the
root cause, the simple-structure model exhibits a definite shift away from the
verbal/communication items. It may be necessary to omit this item from measurement
equivalency studies, or to constrain its loading to the published value. For the other subscales,
only fit, not conceptual-shift, is an issue. For these factors using a more saturated baseline model
or permitting a number of correlated residuals should provide sufficient fit.

It is also possible that scoring discrepancies may impair longitudinal analyses. Latent growth
models and latent growth-mixture models are common longitudinal models. Both of these
models fit random trajectories (i.e., level, slope) to individual scores across occasions. If some
individual scores are biased up or down, both the level and slope growth parameters are biased, as well as their variances. This issue has not been systematically studied and the actual practical consequences are unknown.

In general, one possible approach to avoid both scoring problems and conceptual shifts might be to select smaller item sets for each subscale that are not factorially complex and have high and homogeneous loadings (i.e., exhibit simple structure). This may result in selecting a subset of items that represents a different facet or angle on the concept the scale was originally designed to tap; that is, statistically purer but conceptually limited.

Item subsets with more ideal properties might also benefit efforts to investigate measurement equivalence with multi-group CFA models. A greater degree of simple structure in an abbreviated instrument might obviate the need to use more saturated models or correlated residuals. More importantly, this might allow the use of univariate CFA models which are generally more likely to converge – especially multi-group CFA models. There are many practical difficulties in estimating multi-group models with five factors and 90 items because of the number of model parameters involved.

Limitations of this study include the inability to unambiguously disentangle the sources of incongruence and scoring discrepancies. If the DEKKER sample was reassembled, sampling variation would be eliminated as an alternative explanation and the effect of missing data treatment and statistical methodology on the EFA solution could be better quantified. Similarly, the effects of imposing simple-structure could be unambiguously observed if a simple-structure CFA model was fit to the DEKKER sample. It would be very interesting to see the performance of the item about standing too close (Stands, #78). However, investigation of the ACAD epidemiological sample is not without merit. Longitudinal data have been collected on this
sample, and so it is only in this sample that longitudinal changes in subscale scores are likely to be studied.

Other limitations pertain to the arbitrary nature of comparison criteria employed. For example, discrepancies between the published loadings and our CFA models were highlighted based on an absolute difference of 0.10 or 0.20. This was an arbitrary choice. Also, the gamma statistic used in the scoring tabulations was computed based on the 50th percentile. This was an arbitrary choice, and may not generalize to whatever percentiles represent clinically significant thresholds.

Overall, the framework articulated here and the issues raised should guide the ongoing use of the DBC in individual differences research, and inform the specification of baseline models in any subsequent efforts to study measurement equivalency across age and IQ levels. Further, consideration of the issues raised in this study might benefit other researchers who are trying to employ or adapt inventories originally developed for clinical screening purposes.
References


Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*, (You can get a copy of this paper by emailing bmuthen@ucla.edu. It was never revised and resubmitted).


Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.


Appendix I – Manifest Change vs. True Change

Cattell (1966, p. 366) offered a hypothetical example where shrinking and stretching of the relation between unit change on the manifest scale and unit change on the unobserved true score could, in extreme cases, completely reverse the rank order of individuals. Figure I1 shows such a hypothetical relationship. At the low end of the true score range, the scale score rises more slowly, while at the upper range of true scores, the scale score increased rapidly.

Based on this curve, Table I1 shows observed raw scores and corresponding true score for three individuals at two occasions under two different true score change scenarios. In the upper part of the table, the three individuals all show a uniform true score growth of two units, and non uniform corresponding raw score changes due to the non-linear scaling. Since there is no variance in true score change, the change in observed raw scores and true score change is uncorrelated despite the perfect correlation of the observed row scores with the true scores at both occasions. The rank order of the three individuals is unchanged from time one to time two. This is also true in the lower part of the table where greater true score growth occurs at the lower end of the scale. Individual’s observed raw scores and the true scores have the same rank order at both occasions. However,
now that there is variation across individuals in true growth (i.e., 3,2,1), the correlation of raw score change and true score change is defined. While the correlations across occasions of both the observed raw score and the true score are both 1.0, the correlation of raw change scores with true change scores is -1.0. This is an extreme case, but it underscores the point that the generally high correspondence of raw scores and factor scores (a proxy for the true score) at all occasions, does not guarantee similar correlations between change scores unless the scaling is strictly linear. Figure I2 contrasts the observed raw and true change scores given in Table I1.

Table I1. Observed raw scores, true scores and change scores for three individuals under two patterns of true score growth.

<table>
<thead>
<tr>
<th>True Growth</th>
<th>Raw Scale</th>
<th>True Scale</th>
<th>Raw Score Correlation</th>
<th>True Score Correlation</th>
<th>ΔR,ΔT Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform</td>
<td>T1 T2 ΔR</td>
<td>T1 T2 ΔT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 11 5</td>
<td>2 4 2</td>
<td>1.0</td>
<td>1.0</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>14 20 6</td>
<td>5 7 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 35 12</td>
<td>8 10 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greater At Low-end</td>
<td>5 11 6</td>
<td>1 4 3</td>
<td>1.0</td>
<td>1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>15 23 8</td>
<td>6 8 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 35 10</td>
<td>9 10 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure I2. Observed change magnitude contrasted with true change.
Chapter 3: Evidence for Age and Sex Factorial Invariance of the Eysenck Personality Questionnaire in Later Adulthood

There have been few factorial invariance studies of measurement properties of personality inventories (e.g., Collins & Gleaves, 1998; Hofer, Horn, & Eber, 1997; Quereshi, 2000) and no CFA studies of factorial invariance of general personality measures in later adulthood. An IRT analysis of the revised Eysenck Personality Questionnaire - Short Form (EPQR-S) Extroversion and Neuroticism scales did find some evidence of minor differential item functioning between younger and older adults (Mackinnon et al., 1995). While many studies have investigated the replicability of personality structures across cultures (e.g., P. Barrett & Eysenck, 1984; P. T. Barrett, Petrides, Eysenck, & Eysenck, 1998; Caprara, Barbaranelli, Bermudez, Maslach, & Ruch, 2000; Katigbak, Church, & Akamine, 1996), across age and occasion (e.g., McCrae, Costa, & Arenberg, 1980), and across rater (McCrae et al., 2004); these studies generally used exploratory factor analysis (EFA), and only focused on the consistency (i.e. pattern) of personality structure, not on measurement invariance of the items. Likewise, other work which has modeled the mean and covariance structure of personality dimensions across cross-sectional and longitudinal age (Loehlin & Martin, 2001), relied on scale scores and assumed item level invariance.

The question of measurement comparability is particularly pertinent to gerontological research where developmental changes in health and social behaviors may affect the likelihood of older adults endorsing items designed for younger adults. For example Schneider & Gibbons (1982) hypothesized that declining health might interact with the somatic content of some neuroticism items on the Eysenck Personality Inventory (EPI), or that older adults might not relate to extroversion scale items about party going or playing pranks. Other research with
adolescents and young adults using several forms of the Eysenck Personality Questionnaire has found evidence of modest sex bias in Neuroticism items (Forrest, Lewis, & Shevlin, 2000; Francis, 1993) and Lie items, though there is some evidence that bias in Neuroticism items can be attributed to socialization rather than biological sex (Shevlin, Bailey, & Adamson, 2002). Francis and colleagues attributed bias in the EPQR-S’ Lie scale to a subset of items loading on a social conformity factor in addition to a dissimulation factor, however this bi-dimensionality may only manifest in liberal Christian cultures (Katz & Francis, 1991; Loo, 1995). It is not clear how these findings extend to older adults where gender-typical attributes may differ (Cooper & Gutmann, 1987; Huyck, 1990; McGee & Wells, 1982), or to earlier birth cohorts where gender socialization may have differed (Hyde, Krajnik, & Skuldt-Niederberger, 1991).

Investigations of the measurement properties of personality instruments in populations of older adults can inform comparisons of mean level change in personality traits across age and sex in late life, as well as extend basic theoretical understandings of personality. For example, decreased social support and increased unmet needs in old-old women were found to be associated with higher neuroticism scores, while lower extroversion scores were associated with poorer health and greater psychosocial needs (Maiden, Peterson, Caya, & Hayslip, 2003). If there is no item-bias, and such scale score trends reflect true trait level change, an interactionist framework for personality is indicated and frameworks emphasizing stability should be questioned (McCrae et al., 1980). However, other research has found Neuroticism and Extroversion at the latent trait level to be uncorrelated with age in older adults (Mackinnon et al., 1995), or that while scale scores for Extroversion, Neuroticism, Psychoticism, did decrease with age, age was only a weak-predictor of scale scores and adult heritabilities were stable (Loehlin & Martin, 2001). Where there is across age (or age-sex) factorial invariance permitting conclusions
that true scores are unchanging, theoretical frameworks emphasizing stability of adult personality are advanced.

Item level measurement invariance is a crucial assumption of models of longitudinal change. Studies of individual level personality trajectories require comparable measurement at each occasion. Recently variability in trajectories of extroversion and neuroticism were found to be associated with age-graded life events (Mroczek & Spiro, 2003) and with risk of mortality (Mroczek & Spiro, 2005). Longitudinal comparisons of intra-individuals personality scale scores usually rely on the untested assumption that these concepts have been measured in identical ways and therefore have identical meanings across groups. Over different occasions with the same persons, the items are assumed to measure the same thing. However, with changing life circumstances and physiological and behavioral changes that accompany aging, certain items may be endorsed more, or less, and so contribute to the factor concept (e.g., covariance, mean level) in different ways later in life. In research in adulthood and later adulthood we wish to make comparisons across groups of individuals differing in age, birth year, or time of measurement, as well as across occasions within individuals. Unambiguous comparisons can be made only if the measurement properties of the variables do not differ across groups or occasions. Measurement invariance is a hypothesis that can be empirically tested in a CFA framework.

Factorial invariance provides evidence that the quantity of a construct or factor is proportionally equivalent across groups. By “proportional” we mean that the construct (factor) itself may be higher or lower in level, variability, or covariance with other factors or constructs, but that the factor-item regressions (factor loadings) and perhaps the unique means (intercepts) are the same. If the invariance of factor loadings is not obtained, it is impossible to know
whether group differences in factor means, factor variances, or the relationships of a factor with other constructs are due to differences at the measurement level (e.g., different mixtures of items identifying a factor) or to true differences in the factors themselves. Furthermore, lack of evidence for or against measurement invariance implies that we must regard group differences as preliminary. Until such a test of invariance has been performed, we cannot firmly state that we have measured the constructs in the same way. It is for this reason that evidence for factorial invariance (or lack of it) has important implications for developing measures of personality, for age-based individual differences research, and for their application in comparative research using different samples.

Comparison of mean structure across groups is also fundamental to tests of personality structure and is a major outcome of factorial invariance across groups. Of what use is an assessment instrument if it does not accurately indicate differences in level of personality characteristics in one person or group of individuals compared to another? Nevertheless, most recent work on the structure of personality and cognitive ability factors largely ignores tests of the mean structures, an essential aspect of Meredith’s (1993) hierarchy of invariance tests (described below). While the expectation is that factor level means would differ across groups, the unique means (intercepts) of the factor indicators should be invariant such that the combination of item-level averages are comparable at the factor level. Multiple-group (e.g., across age-cohorts) models of factor invariance will permit strong tests, based on the reliable variance, of mean differences across age or sex groups at the factor level.

The Present Study

The purpose of this study was to investigate the factor structure of the Extroversion, Neuroticism, and Psychoticism scales of the Eysenck Personality Questionnaire – Short Form
(EPQR-S) in order to evaluate evidence for its factorial invariance across sex and age groups in later adulthood. Multi-group polytomous CFA models will be used for testing invariance hypotheses. Previous investigations (Forrest et al., 2000; Shevlin et al., 2002) of sex bias in the EPQR-S and EPQR-Abbreviated (EPQR-A) employed MIMIC models (Finch, 2005; Rubio, Berg Weger, Tebb, & Rauch, 2003) which can test intercept invariance, but risk mis-specification because such models assume across-groups invariance of loadings and common-factor variance. Based on the level of factorial invariance obtained, age and sex-based differences in factor-level means and inter-factor association will be assessed.

**Method**

**Measures**

Eysenck Personality Questionnaire – Short Form (EPQR-S). A 48-item short scale version (H. J. Eysenck & Eysenck, 1994; S. B. Eysenck, Eysenck, & Barrett, 1985) of the 100 item EPQ (H. J. Eysenck & Eysenck, 1975) consists of 12 items for each of three EPQ-R higher-order dimensions of personality: Extraversion (E), Neuroticism (N), and Psychoticism (P); plus an additional 12-item Social Conformity or Lie Scale. (Full text of items is listed in Appendix I). Each item has true/false response format (1=true; 0=false). Items are reverse scored when necessary.

**Subjects**

The sample was a subset of twins from a volunteer population-based sample ascertained for research on tobacco and alcohol use. The original sample pool was restricted to Caucasian twins by design (see Meyer, Heath, & Eaves, 1992 for details on sample ascertainment). Subjects were adult twins 50-96 years of age, originally recruited between 1985 and 1989 through a
newsletter published by the American Association of Retired Persons (AARP); hereafter we refer to our sample as the AARP sample. If either twin responded to this original recruitment, both pair members were mailed an initial questionnaire that included the 48 item EPQR-S (H. J. Eysenck & Eysenck, 1994; S. B. Eysenck et al., 1985) in addition to several other personality inventories.

At the time of original ascertainment, 72% of the sample was married or cohabitating, with males more likely to be married or cohabitating (88%) than females (67%). In addition, respondents were relatively well educated, with 52% of the females and 65% of the males having some education beyond high school. The sample also tended to be generally of middle class SES, with approximately 43% reporting an annual income of at least $35,000 (For a more detailed description of this sample, see Prescott et al., 1994).

Several exclusion criteria were applied to the data. Valid questionnaires were obtained from 4119 subjects (75% female), representing a response rate of 63.5% (approximately 40% of the originally ascertained sample). All EPQR-S items were missing for 93 respondents, and nine additional respondents failed to respond to at least eight of the twelve items for at least one of the E, N or P subscales. For six of these, only the last 12 items had a response, which suggests an issue with questionnaire layout. Nineteen additional respondents did not respond to the last 12 items, which can also be attributed to layout issues. Although these 19 respondents were disproportionately older, the missed items were proportionately distributed across each scale, so these 19 cases were retained. The retained sample (N=4017 individuals in 2604 twin pairs, 1413 where both twins provided data) consisted of 2990 females (74.4%) and 1027 males (25.6%). Males were slightly older than females ($M_F = 66.5; SD = 7.87; M_M = 67.5; SD = 7.90; t = -3.52, p < .001), and the proportion of females increased with age. The preponderance of females
overall and with age probably reflects both the earlier mortality of men and the higher tendency of females to participate in volunteer studies.

Just under 12% of participants (F=361, M=109) scored 11 or 12 out of 12 on the Lie Scale which raises issues of validity. However, past research which manipulated the motivation for dissimulation, provided evidence that in addition to dissimulation, the Lie Scale also measures a stable personality trait that might be interpreted as social naivety or conformity (H. J. Eysenck & Eysenck, 1994).

Due to the need for large samples when working with dichotomous items, both twins were retained if available. The dependency due to the nesting of individuals within twinships can potentially shrink standard errors and may even distort covariance structure. However this must be considered in the light of the potential for substantial bias in the polychoric correlation estimates when some response options have a small absolute number of endorsements (Brown & Benedetti, 1977). Some of the EPQR-S items tend to be infrequently endorsed – especially on the Psychoticism Scale. (For example, “Do you want people to be afraid of you?”) The approach used here will be to retain all respondents and to check for any consequential bias.

Three age groups (50-62, 63-69, 70-96) were defined to obtain relatively equal proportions of participants in each age group (with slightly more in the older groups to offset increasing patterns of item missingness observed with age). In the youngest group, there were 1243 (77% female), and 1368 (75% female) and 1406 (71% female) respectively in the two older groups.

Methodological Considerations

Sibling Dependencies. To investigate the effects of sibling dependencies, four-factor EFA models can be estimated using both all available data and a single-sibling subset. The pattern and magnitude of factor loadings in each model can be compared to evaluate potential distortion of
the structure. Also, the inter-factor correlations from each model can be compared for additional information. These approaches are both exploratory.

*Empty-cells in bivariate cross-tabulations.* The polychoric correlation of dichotomous items is undefined (i.e., asymptotically goes to ±1.0) when the cross-tabulation contains a zero-cell, and substantial bias may be introduced when cell counts are very small (Brown & Benedetti, 1977; Divgi, 1979; B. Muthen, 1989). In practice, small non-zero constants can be placed into zero cells when zero is deemed to be a poor approximation of the true population value, but the estimate obtained is very sensitive to the constant used (K. G. Joereskog, 2002). Also, Brown and Benedetti stress that there is increasing potential for bias as cell counts drop below five, (see graphical illustration in Chapter 1). These conditions are more likely when items with highly skewed distributions have high pairwise correlations.

These issues are compounded in multi-group and multi-factor models. Even low cell counts are likely to result in empty-cells when the data are divided into multiple groups. This results in increased potential for distorted factor structure in one or more groups in the unconstrained baseline models. In multifactor models, unfortunate bivariate cross-tabulations can occur not only within the items loading on individual factors, but also between items loading on separate factors, with the potential to distort the inter-factor correlations. This situation is most likely with factorially complex items that have cross-loadings not represented in simple structure models. When sufficient data to avoid multiple zero-cells or cells with low counts are not available, problematic items, or even scales, should be excluded from analyses.

*Evaluation of Factorial Invariance.* Multi-group confirmatory factor analysis (M-CFA) assessment of factorial invariance provides evidence for or against the comparability of personality factors across age and sex groups in late adulthood. MCFA models permit testing of
all the parameters of the factor analytic measurement model relating items to latent common-factors (i.e., loadings, intercepts, unique variances), Procedures for traditional CFA models are well established (Hofer et al., 1997; Horn & McArdle, 1992; Meredith, 1993; Widaman & Reise, 1997), although not without controversies of nomenclature and recommended procedures (Bontempo & Hofer, in press; Vandenberg, 2002; Vandenberg & Lance, 2000). Only recently, following software development, have procedures for testing factorial invariance CFA models with polytomous indicators received discussion (Millsap & Yun-Tein, 2004; B. O. Muthen & Asparouhov, 2002). Prior studies of bias in polytomous CFA models had use software for continuous indicators, or procedures using EFA and rotation.

Evidence for measurement invariance involves a multi-group baseline CFA model and a nested sequence of increasingly constrained CFA models (invariance hypotheses) based on Meredith’s (1993) formulation of strong and strict factorial invariance. Model fit is judged by one or more goodness of fit indices (GFIs, Bentler, 1990; Browne & Cudeck, 1993; Cheung & Rensvold, 2002). Meredith defines metric invariance where loadings are equivalent across groups, and scalar invariance when intercepts are invariant across groups. An invariance hypothesis is retained when a nested model (where model constraints are a subset of another less constrained model) imposing across-group equivalence constraints does not significantly increase misfit when compared to the fit of the baseline model (i.e., small change in goodness of fit indices, ΔGFI). A model judged to fit well provides some support for invariance hypotheses, but ΔGFI can provide a test that is independent of the fit of the baseline model (Cheung & Rensvold, 2002). Strong factorial invariance obtains when hypotheses of metric and scalar invariance are both supported. Strict invariance obtains when the hypothesis of invariant unique variances is also supported.
There is an established sequence of nested models beginning with the baseline model and additively testing hypotheses of weak, strong, and strict factorial invariance (Cheung & Rensvold, 2002; Hofer, 1999; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000; Widaman & Reise, 1997). The specific measurement parameter constraints are described below and correspond to parameters shown in Figure 1.

1) *Configural invariance.* This logical hierarchy of constraints begins with Thurstone’s concept of *configural invariance* (Thurstone, 1947) (see also Horn & McArdle, 1992; Reise et al., 1993), which requires that only the same number of factors and pattern of salient factor loadings be equivalent across groups. This model tests the hypothesis of the zero-loadings needed to specify some degree of simple structure. Using this model as a baseline, the decrement in fit associated with more constrained nested models can be determined.

2) *Weak factorial invariance* involves imposing equality constraints across groups on the factor loadings or factor-variable regressions (shown as $a_2$, $a_3$, $a_4$, etc. in Figure 1) while ensuring the factor variances and covariances are free to vary. Note the factor variance or one of the loadings is fixed (in this case, to 1.0 as shown in Figure 1) to identify the metric of the factor. That the factor loadings are found to be invariant is not to say that they are identical since the factor variances and covariances must be free to vary across groups. Rather, we say that the factor loadings (or regression weights) in one group are said to be *proportionally equivalent* to corresponding loadings in other groups. Weak factorial invariance must be established in order to allow unambiguous comparison of groups with respect to common-factor variances – but not means.
3) **Strong factorial invariance** permits comparison of factor means across groups. This test requires the additional constraint of invariant mean intercepts across groups which forces mean differences across groups to be expressed at the factor level. The mean intercepts are shown in the graph as a triangle (mean intercepts to each manifest variable are not shown). Because of the way constraints must be placed on the model, factor mean differences across groups are expressed as differences relative to an arbitrary group [e.g., factor means are set to zero, as shown in Age Group 1 in the figure, or some other arbitrary value (Sorbom, 1974; see also, Horn & McArdle, 1992; Meredith & Horn, 2001) ].

4) **Strict factorial invariance** requires the additional constraint of invariant corresponding unique variances ($e_1$, $e_2$, $e_3$, etc.) as well as unique means and factor loadings. This model forces the combined specific and random error components of each variable to be equivalent across groups such that differences in variance across groups are permitted only at the latent variable level. If a strict invariance model can be fit to the data, we can be well assured that measurement comparisons across samples involving factor mean and factor covariance structures are valid. In addition, strict factorial invariance is necessary for valid comparisons based on unit-weighted sums of scale items.
Polytomous Factorial Invariance. Multi-group CFA measurement models with binary indicators (i.e., the yes/no response format for the EPQR-S items) must be parameterized differently (Karl G. Joereskog & Moustaki, 2001; Millsap & Yun-Tein, 2004; B. O. Muthen & Asparouhov, 2002), and this necessitates modifications to the procedure described above. For each manifest indicator, these models utilize an underlying latent continuous response variable cut by \( m-1 \) threshold parameters (where \( m \) is the number of response options) so as to produce the observed response frequencies. Analysis is then based on the tetrachoric correlations of item thresholds. Since these threshold parameters are what connect the measurement model to the actual data, their invariance is an important issue.

The additional measurement model parameters (i.e., thresholds, latent response variable’s mean and variance) present issues for both model identification and factorial invariance.
procedures. The use of latent response variables brings the analysis into correlation metric, and additional scaling factors are needed to capture across-group differences in the common factor mean and variance. To identify the model, the intercept parameters for all latent response variables are fixed to zero, and their uniqueness variances are fixed to one in the first group. This requires that a test for strict invariance will involve fixing the uniqueness parameters to 1.0 in all groups (as opposed to the usual across-group equivalence constraints) because they are fixed in the first group. Also, as with traditional CFA multi-group models, further constraints are needed to place the common-factor mean and variance into comparable across-group metrics.

The additional constraints can be placed on either thresholds or uniqueness parameters. Millsap and Tein (2004) work out minimally necessary across-group threshold constraints to identify the model and permit multi-group model testing. They require that the first of the m-1 thresholds be constrained across all groups, and a second threshold (or a uniqueness when there is no 2nd threshold) to be constrained for one reference item in each group. For dichotomous items these identification choices do not permit the hypothesis of threshold invariance to be tested against a less constrained baseline model, but a well-fitting baseline model can be taken to constitute some evidence for threshold invariance. Muthen and Asparouhov (2002) use a different approach where thresholds and loadings are constrained in a reduced model and tests of selected items are conducted against a full model where the threshold and loading for these items are freed while maintaining model identification by fixing the specific-variance to unity for the selected items (see Muthen and Asparouhov’s footnote 3 on p. 11). Items to be tested might be selected by examining non-response and endorsement rates across groups, through the use of MIMIC models that are sensitive to threshold invariance, and/or by reliance on modification indices. This is a departure from the hierarchical least-constrained to most-constrained
procedures outlined earlier\textsuperscript{a}. However, it is also possible to use the uniquenesses instead of thresholds to identify a minimally constrained baseline model and still work in a hierarchical fashion.

Given the need to constrain either thresholds or uniquenesses to identify multi-group polytomous CFA models, it is not possible to test all measurement parameters in a single hierarchical sequence. In either case, the same number of parameters are constrained or fixed. Consider a two group factor model with four dichotomous indicators for a single-common factor. In total there are 18 potential measurement parameters in each group (see Table 1): four thresholds, four uniqueness, four intercepts, four loadings, one common factor mean, and one common-factor variance. In all multi-group CFA models there must be a reference group (Reise et al., 1993), where a variance parameter (i.e., loading or common-factor variance) and a mean-structure parameter (i.e., intercept or common-factor mean) are fixed to provide a scale for the latent factor, so there are two fewer parameters estimated in the reference group. In both baseline models considered here, the intercept parameters are fixed to zero, the loadings are freely estimated, and the common-factor variance is fixed to one in all groups since the loadings are freely estimated in each group (i.e., the factor variance is not yet on a common metric across groups).

\textsuperscript{a} Beginning with the constrained model and freeing up selected parameters is consistent with the concept of partial measurement invariance (Byrne, Shavelson, & Muthen, 1989).
The difference in baseline models then, involves only the threshold parameters in all groups, the uniqueness parameters in the non-reference groups, and the common-factor mean in the non-reference groups. Also, recall that in both baseline models, the uniqueness parameters must be fixed to 1.0 in the reference group. So in the threshold baseline, one threshold for each item is constrained to be equivalent across groups, the uniqueness for a single item is fixed to 1.0 in both groups, and the common-factor mean is now estimated in the second group as it is identified via the across-group threshold constraints. In the uniqueness baseline, the uniqueness parameter for each item is fixed to 1.0 in all groups, and the common factor means are fixed to zero in each group.

<table>
<thead>
<tr>
<th>Parameter Name</th>
<th>Threshold Baseline</th>
<th>Uniqueness Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Status</td>
<td>#</td>
</tr>
<tr>
<td>threshold (τ₁)</td>
<td>constrained 1</td>
<td>free 1</td>
</tr>
<tr>
<td>threshold (τ₂)</td>
<td>constrained 2</td>
<td>free 2</td>
</tr>
<tr>
<td>threshold (τ₃)</td>
<td>constrained 3</td>
<td>free 3</td>
</tr>
<tr>
<td>threshold (τ₄)</td>
<td>constrained 4</td>
<td>free 4</td>
</tr>
<tr>
<td>intercept (ν₁)</td>
<td>fixed=0</td>
<td>fixed=0</td>
</tr>
<tr>
<td>intercept (ν₂)</td>
<td>fixed=0</td>
<td>fixed=0</td>
</tr>
<tr>
<td>intercept (ν₃)</td>
<td>fixed=0</td>
<td>fixed=0</td>
</tr>
<tr>
<td>intercept (ν₄)</td>
<td>fixed=0</td>
<td>fixed=0</td>
</tr>
<tr>
<td>loading (λ₁)</td>
<td>free 5</td>
<td>free 5</td>
</tr>
<tr>
<td>loading (λ₂)</td>
<td>free 6</td>
<td>free 6</td>
</tr>
<tr>
<td>loading (λ₃)</td>
<td>free 7</td>
<td>free 7</td>
</tr>
<tr>
<td>loading (λ₄)</td>
<td>free 8</td>
<td>free 8</td>
</tr>
<tr>
<td>uniqueness (θ₁)</td>
<td>fixed=1</td>
<td>fixed=1</td>
</tr>
<tr>
<td>uniqueness (θ₂)</td>
<td>fixed=1</td>
<td>fixed=1</td>
</tr>
<tr>
<td>uniqueness (θ₃)</td>
<td>fixed=1</td>
<td>fixed=1</td>
</tr>
<tr>
<td>uniqueness (θ₄)</td>
<td>fixed=1</td>
<td>fixed=1</td>
</tr>
<tr>
<td>factor variance (ψ)</td>
<td>fixed=1</td>
<td>fixed=1</td>
</tr>
<tr>
<td>factor mean (α)</td>
<td>fixed=0</td>
<td>fixed=0</td>
</tr>
</tbody>
</table>

Table 1. Estimated parameters in threshold and uniqueness baselines.
because all thresholds are estimated. In this example there are 16 distinct parameters estimated regardless of which baseline is used.

The approach in this study is to use both baseline models. Either baseline can be further constrained to specify a pattern of measurement parameters that represents strict factorial invariance, (i.e., constrained thresholds, loadings, and uniquenesses), but neither baseline can provide a true test of strict factorial invariance because each baseline must assume either threshold or uniqueness invariance. First, using the uniqueness baseline, hypotheses about invariant loadings and thresholds will be tested. This follows the sequence used for traditional CFA models. Second, the hypothesis of uniqueness invariance will be tested against a reduced version of the threshold baseline in which loadings will also be constrained. Ideally, all invariance hypotheses would be supported. However, should either threshold or uniqueness invariance hypotheses not be supported, any other invariance hypotheses tests bases on such an assumption might be biased. One implication of this confound is that unlike traditional CFA invariance procedures, where the test of strict factorial invariance might be omitted if only the common-factor mean and variance is of interest, in polytomous CFA procedures, the test of uniqueness invariance should not be omitted because it is interdependent with the tests of threshold and loading invariance.

*Multi-factor models and simple-structure.* There is some disagreement as to whether factorial invariance tests of inventories with multiple common factors should be conducted with multi-factor or single-factor models. In larger models, exceptional fit in one part of the model can mask problematic misfit in another part of the model. In this sense a single-factor model provides a more sensitive test. However, the pattern of zero cross-loadings can only be tested in multi-factor models. Also, different patterns of salient loadings may emerge in multi-factor
models with some degree of simple structure when the common factors have non-negligible correlations. This occurs because the item variance is explained not only by the direct effect of its common factor, but also by the indirect effect of correlated factors. When there is misfit because non-negligible cross-loadings are omitted, the common-factor correlations (and consequently item loadings) may shift to absorb the effects of mis-specification. In extreme cases there could be poor correspondence of the item loadings in single-factor and multi-factor models, and invariance obtained in one model may not generalize to the other.

Even when all important cross-loadings are included in multi-factor models, poor correspondence to single-factor solutions can still result when there are sizeable factor correlations, because the single-factor model implicitly holds factor correlations at zero. In this case, estimates for common-factor variance and item-loadings may differ because item variance is no longer partly explained by the indirect effect of the correlated factors. Ultimately the degree of non-congruence that can be tolerated will depend on the breadth of the factor and the amount of implicit aggregation in the items. For example an EPQR-S item such as “Are you rather lively?” asks the respondent to aggregate over more occasions and contexts than an item such as “Can you easily get some life into a dull party?” (For further discussion of construct breadth, aggregation, and the sensitivity of factorial invariance tests, see Bagozzi & Edwards, 1998; Bandalos, 2002; Vandenberge, 2002)

In personality research, a well-fitting baseline model often requires estimation of cross loadings and of the covariances between the uniquenesses of some items, and this constitutes a departure from the simple-structure in traditional CFA studies. In many factorial invariance studies of inventories with multiple common factors, the configural invariance evidenced by the good fit of the baseline model is also a test of simple structure, or more specifically a test of the
pattern of zero constraints on cross-loadings and of the mutual orthogonality of the unique variances. The CFA literature advises researchers to be cautious when making post hoc modifications to their models in order to avoid overfitting the data and limiting generalizability.

Unfortunately, previous CFA studies of EFA-derived personality structures have found simple structure models to be very poor explanations of personality inventory data (Church & Burke, 1994; Katigbak et al., 1996; McCrae, Zonderman, Costa, & Bond, 1996; Raykov, 1998). It has been argued that only one non-zero loading for each observed item is not a plausible model in factor-analytic personality research because the factors are construed as superfactors that refer to a diverse set of lower-level constructs such as primary factors, habits, and specific behaviors (Borkenau & Ostendorf, 1990). Similarly, the factor-analytic model asserts mutually orthogonal common and specific factors, but the higher-order nature of personality factors can result in large correlations among specific factors (uniquenesses) due to excluded lower-order factors and/or method variance. This is also consistent with Meredith’s (1993; Meredith & Horn, 2001) demonstration that unique factors which are uncorrelated in the population may not be uncorrelated in selected subpopulations. Recent CFA studies of personality structure have emphasized the utility of CFA in personality research, provided that theoretically justified model modifications are made to obtain good fit (Aluja, Garcia, & Garcia, 2004; Aluja, Garcia, Garcia, & Seisdedos, 2005). This advice is also consistent with recommendations in the factorial invariance literature that studies of instruments with factorially complex items may need to utilize more saturated structures instead of simple structure (Hofer, 1999; McArdle & Cattell, 1994; Meredith & Horn, 2001).

The approach taken here will be to use multi-factor models that have sufficiently saturated structure. This will permit the examination of common factor means and the estimation
of inter-factor correlations across age and sex groups. In general, the goal of the study should
dictate the use of single-factor vs. multi-factor models and simple vs. more-saturated structure. If
the intent is to investigate properties of the common factor means, (e.g., cross-sectional age
trends, gender differences) or associations among common-factors (e.g., covariance of
extroversion and neuroticism in males and females), then multi-factor models will be needed.
Also, more saturated structures might be recommended to avoid distortion of the inter-factor
correlations that might be introduced by imposing a simple-structure. On the other hand, the goal
may be to demonstrate strict factorial invariance in order to validate the use of manifest summary
scores (e.g., a simple sum of scale items) for making quantitative comparisons. In this case,
single-factor models without the inclusion of cross-loading items from other scales should be
used because this is the measurement model implied by manifest summary scores.

_Evaluation of Model Fit._ Our models were estimated using MPlus v4.0 (B. O. Muthen &
Muthén, 1998-2006) using the WLSMV estimator with missing data handled by maximum-
likelihood estimation assuming missing at random. WLSMV provides weighted least squares
parameter estimates using a diagonal weight matrix with standard errors and mean- and variance
adjusted chi-squared test statistic that use a full weight matrix (B. O. Muthen, du Toit, & Spisic,
1997). For nested models the difference between the fit functions is not distributed as chi-square
and the degrees of freedom is not the simple difference in free parameters. However, an
appropriate chi-square statistic and degrees of freedom are calculated according to formulas in
the MPlus Technical Appendicies (at [www.statmodel.com](http://www.statmodel.com)). In any sample of sufficient size,
minor deviations of the observed data from the model-based expectations will lead to statistically
significant chi-square statistics. This is also true of the adjusted chi-square statistic provided for
the WLSMV estimator. Therefore, in addition to chi-square difference tests, several relative fit
indices were used to evaluate overall model fit. The adequacy of model fit was assessed by the MPlus chi-squared difference test for the WLSMV estimator (Satorra & Bentler, 2001), the Comparative Fit Index (CFI, Bentler, 1990), the Tucker-Lewis Index (TLI, Tucker & Lewis, 1973) and the Root Mean Square Error of Approximation (RMSEA, Browne & Cudeck, 1993). The cutoff values TLI>.95, CFI>.95, and RMSEA<.06 recommended by Hu and Bentler (1999) for CFA models with continuous indicators have performed well in simulation studies of CFA models with dichotomous indicators (Yu, 2002). Finally, Chung and Rensvold (2002) have investigated cutoff criteria for change in relative fit indices. Delta-CFI ($\Delta$CFI) was found to be independent of CFI, and to have the best performance in simulations varying model and sample size. They recommend $\Delta$CFI>.01 as the cutoff for rejecting the hypothesis that there is no significant increase in misfit.

Results

The results are sequenced as 1) descriptive statistics, including age-sex group comparisons, 2) preliminary analyses to investigate potential bias from sibling dependencies or low-cell counts, 3) multi-group polytomous CFA models corresponding to invariance hypotheses, and 4) common-factor mean trends across age and sex.

Descriptive Statistics: Age-Sex Group Comparisons

In general, mean scale scores for E, N, P, and L evidenced the expected sex and age patterns, although E and L scores were higher than comparably aged adults in other studies. Table 2 shows mean age, mean scale scores, and scale reliabilities for each of the six age-sex groups in our sample. Comparisons with studies offering similar data are also tabulated. Consistent with previous research that found small age-related decreases in Extroversion, Neuroticism, and Psychoticism, and an increase in dissimulation (Lie) (S. B. Eysenck et al.,
1985; Loehlin & Martin, 2001), the older adults in our study scored lower on E, N and P scales and higher on the L scale than the published adult norms for males and females (H. J. Eysenck & Eysenck, 1994). Also, consistent with the adult sex norms, the mean score for females was higher than the mean score for males on Neuroticism (MF = 4.5, MM = 3.6, \( t = 6.31, p < .001 \)) and this pattern held across all three age groups (MF = 4.7, 4.5, 4.3 vs MM = 3.6, 3.4, 3.8). Likewise for dissimulation the mean score for females was higher (MF = 7.3, MM = 6.7, \( t = 5.78, p < .001 \)) and this pattern held across all three age groups (MF = 7.0, 7.3, 7.7 vs MM = 6.0, 6.5, 7.2). Again, consistent with the adult sex norms, the mean Psychoticism score was lower for females (MF = 1.2, MM = 1.4, \( t = -4.63, p < .001 \)), and this pattern held across all three age groups (MF = 1.1, 1.1, 1.2, vs MM = 1.7, 1.3, 1.3). However for Extroversion, the mean scores for females and males were not significantly different (MF = 7.4, MM = 7.3, \( t = 0.62, ns \)). This pattern also held across age groups (MF = 7.5, 7.5, 7.1 vs MM = 7.0, 7.5, 7.3).

Table 2. Age, mean EPQR-S scale scores, and reliabilities by gender-age, compared with past studies.

<table>
<thead>
<tr>
<th>Scale</th>
<th>EPQR Manuala</th>
<th>Current AARP Study</th>
<th>Selected Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F  M</td>
<td>All Ages F  M Age50-62 F  M Age63-69 F  M Age70-96 F  M</td>
<td>Study 1b F  M Study 2c F  M Study 3d F  M</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Std. Dev.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extroversion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>8.1 7.9</td>
<td>7.4 7.3 7.5 7.0 7.5 7.5 7.1 7.3</td>
<td>5.9 5.8 6.9 4.4 6.6 6.4</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>3.3 3.3</td>
<td>3.4 3.4 3.5 3.6 3.4 3.4 3.3 3.3</td>
<td>3.5 3.6 3.4 3.5</td>
</tr>
<tr>
<td>Reliability</td>
<td>.84 .84</td>
<td>.85 .85 .86 .87 .85 .84 .84 .84</td>
<td>.84 .88</td>
</tr>
<tr>
<td>Neuroticism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>6.7 4.7</td>
<td>4.5 3.6* 4.7 3.6 4.5 3.4 4.3 3.8</td>
<td>3.6 2.7 5.1 4.3 4.2 3.4</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>3.4 3.4</td>
<td>3.2 3.1 3.1 3.2 3.2 3.1 3.2 3.1</td>
<td>3.4 3.4 3.2 3.1</td>
</tr>
<tr>
<td>Reliability</td>
<td>.83 .84</td>
<td>.81 .83 .80 .84 .83 .83 .81 .82</td>
<td>.80 .84</td>
</tr>
<tr>
<td>Psychoticism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.1 2.8</td>
<td>1.2 1.4* 1.1 1.7 1.1 1.3 1.2 1.3</td>
<td>1.7 2.3 1.6 1.9</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>1.9 2.2</td>
<td>1.2 1.3 1.2 1.5 1.2 1.2 1.2 1.3</td>
<td>1.7 1.8 1.3 1.5</td>
</tr>
<tr>
<td>Reliability</td>
<td>.57 .59</td>
<td>.38 .41 .42 .46 .39 .35 .34 .37</td>
<td>.56 .65</td>
</tr>
<tr>
<td>Lie</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>5.3 4.5</td>
<td>7.3 6.7* 7.0 6.0 7.3 6.5 7.7 7.2</td>
<td>5.7 5.8 5.8 5.0</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>3.2 3.0</td>
<td>2.5 2.8 2.6 2.9 2.5 2.8 2.4 2.6</td>
<td>2.5 2.6 2.5 2.5</td>
</tr>
<tr>
<td>Reliability</td>
<td>.74 .74</td>
<td>.70 .75 .71 .77 .69 .75 .68 .70</td>
<td>.73 .77</td>
</tr>
</tbody>
</table>

* Norms for USA EPQR-S in the Manual of the Eysenck Personality Questionnaire (HJ Eysenck, & Eysenck, 1994).
* Canberra Longitudinal Study of Aging data (Mackinnon, et al., 1995). Ages 70-80+ (M=76.28, SD=4.74).
* 51-60 and 61-70 age strata in revised EPQ validation samples (SB Eysenck, Eysenck, & Barrett, 1985).
* Oldest cross-sectional sample in Loehlin & Martin (2001)
* Reliability analysis in our AARP study is based on subsets (85% - 96%) with list-wise complete data for each scale.
* T-test significant at \( p < .001 \).
Compared to British (S. B. Eysenck et al., 1985; Loehlin & Martin, 2001) and Australian (Mackinnon et al., 1995) samples of comparably aged adults, the older adults in this study have higher mean scores on Extroversion, lower mean scores on Psychoticism, and considerably higher mean scores on the Lie Scale. The mean Neuroticism scores for older adults in other studies varied considerably, however the mean Neuroticism scores for the older adults in this study fall within the range seen in these other studies. It is not clear whether the origin of these discrepancies is due to true differences in the American population or sampling biases arising from self-selection into our sample. The high dissimulation scores are of concern.

**Dissimulation.** Due to the higher than expected Lie-scale scores, we next examined mean values of N and P scale scores for a high-lie and a low-lie group, as well as examined the association of the L-score with both N- and P-scores. There is some evidence that the measurement of Neuroticism is vulnerable to faking (McKelvie, 2004). The EPQR manual (H. J. Eysenck & Eysenck, 1994) recommends creating groups at the 95th percentile of the L-scores, and excluding the participants in the high-lie group if the respective N- and P-scale means and L-N or L-P correlations show large differences, but they do not suggest a particular threshold. We created groups by placing participants scoring 10 or 11 into the high-lie group (F = 377, M = 131) and the remaining participants in the low-lie group (F = 1378, M = 658). Because of the restricted range of L scores and the small number of males in the high-lie group, we did not compute the correlation of L with P and N scores in each group as suggested in the manual. Instead we assessed the overall association of L with P and N scores for males and females without grouping. Table 3 shows the N and P means by sex and high- vs low-lie grouping. Also shown is the $R^2$ for the respective regressions of N and P scale scores on L scale scores.
After examining associations with high L-scores, it was decided to retain the high-lie group in the main sample for analysis. The N-scale mean scores for high-lie participants were significantly lower for both females ($M_{LL} = 4.74$, $M_{HL} = 3.48$, $t = 7.45$, $p < .001$) and males ($M_{LL} = 3.79$, $M_{HL} = 2.74$, $t = 4.05$, $p < .001$) but not significantly different for either females or males on the P-scale. For both N and P scales, the amount of variance predicted by L-scores was small. For female N scores only 2.4% of the variance was predicted by L-scores, and for male N scores only 1.6% of the variance was predicted. For P scores, only about 0.5% was predicted.

Since any dissimulation effect seemed negligible, we retained all respondents regardless of Lie scale score. However, we decided to include the manifest Lie scale score as a covariate in the invariance models of the EPQR-S personality factors. In addition to measuring dissimulation, the L scale also measures a stable personality trait similar to social conformity. Possibly higher social conformity in older participants is a difference between American and Australian or British populations.

**Item Endorsement and Non-response Rates.** The missing-data rate and item endorsement rate for each item is tabulated in Table 4. Different rates of item endorsement across age-sex groups for only a subset of a scale’s 12 items is an indication of potentially biased items. Also, systematic differences in non-response patterns across age-sex groups may indicate items whose psychometric properties are problematic. In one study that looked at incomplete response

### Table 3. Contrast of P and N means in high vs low Lie-score groups

<table>
<thead>
<tr>
<th></th>
<th>L Mean</th>
<th>N Mean</th>
<th>$N R^2$</th>
<th>P Mean</th>
<th>$P R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Lie</td>
<td>10.43</td>
<td>3.48*</td>
<td>2.36*</td>
<td>1.13</td>
<td>0.55*</td>
</tr>
<tr>
<td>Low Lie</td>
<td>6.50</td>
<td>4.74</td>
<td></td>
<td>1.16</td>
<td></td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Lie</td>
<td>10.53</td>
<td>2.74*</td>
<td>1.61*</td>
<td>1.30</td>
<td>0.44</td>
</tr>
<tr>
<td>Low Lie</td>
<td>5.91</td>
<td>3.79</td>
<td></td>
<td>1.43</td>
<td></td>
</tr>
</tbody>
</table>

* Regression weight significant at $p<.001$

* T-test significant at $p<0.001$
patterns in an elderly sample (Mackinnon et al., 1995), the average omission rate was 2%,
(varying from 0.9-5.6%), the omission rate was unrelated to position in the inventory, and IRT
measurement parameters were unaffected by including cases missing one to four items per scale;
however the relationship between N and E scale scores and the corresponding latent trait scores
differed significantly between complete responders and partial responders. There is some
correspondence ($r = .48$, root-mean-square-difference RMSD = 20.96) between endorsement
rates for Extroversion items in our AARP data and the rates for comparably aged adults reported
by Mackinnon and colleagues. The correspondence of Neuroticism endorsement rates was much
higher ($r = .94$, RMSD = 10.75). In our AARP data the Extroversion items had the highest mean
non-response rate, followed by Neuroticism, Lie, and Psychoticism items.

The mean rate of endorsement was highest for Extroversion items 61%, while mean rates for
Neuroticism items (35%) and Psychoticism items (10%) were considerably lower. Although item
endorsement rates for both the Extroversion and Neuroticism scales are elevated compared to
rates reported by Mackinnon and colleagues (Mackinnon et al., 1995), the ratio is approximately
two to one in both studies, and the elevated rates are consistent with the higher mean
Extroversion and Neuroticism Scale scores in our AARP data (see Table 1).

Multiple items on the Extroversion, Neuroticism, and Psychoticism Scales were
disproportionately endorsed ($\chi^2$ significant at $p < .05$) across sex or age grouping. On the
Extroversion Scale, items 3, 15, 23, 32, and 41 were disproportionately endorsed by females and
males, while 3, 15, 32, 41, 44, and 48 were disproportionately endorsed across two or more age
groupings. (These tests are presented for descriptive purposes and no control is made for Type I
error.) On the Neuroticism Scale almost every item (5, 9, 13, 21, 25, 36, 38, 42, 46) was
disproportionately endorsed across sex, while only two items (13, 25) were disproportionately
endorsed across one or more age groupings. On the Psychoticism Scale, five items (2, 6, 14, 28, 39) were disproportionately endorsed across sex, and three items (6, 14, 39) were disproportionately endorsed across one or more age groupings. In some cases items were disproportionate across both sex and age groupings, and this may indicate an age-sex item bias. Particular attention can be given to these items if invariance across sex or age does not obtain.

Table 4. EPQR-S Item endorsement rates by age-gender, total, and selected studies.

<table>
<thead>
<tr>
<th>EPQR-S Item #</th>
<th>Short Text</th>
<th>% Missing</th>
<th>Age 50-62</th>
<th>Age 63-69</th>
<th>Age 70-96</th>
<th>Total</th>
<th>Canberra*</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-03</td>
<td>Are talkative</td>
<td>1.4</td>
<td>13.5</td>
<td>4.1</td>
<td>13.8</td>
<td>5.5</td>
<td>6.2</td>
</tr>
<tr>
<td>E-07</td>
<td>Are lively</td>
<td>1.1</td>
<td>17.1</td>
<td>6.2</td>
<td>18.8</td>
<td>7.8</td>
<td>18.1</td>
</tr>
<tr>
<td>E-11</td>
<td>Enjoy meeting people</td>
<td>1.0</td>
<td>19.7</td>
<td>6.9</td>
<td>20.6</td>
<td>9.4</td>
<td>21.2</td>
</tr>
<tr>
<td>E-15</td>
<td>Let go and enjoy party</td>
<td>1.1</td>
<td>13.4</td>
<td>5.5</td>
<td>15.5</td>
<td>7.4</td>
<td>14.7</td>
</tr>
<tr>
<td>E-19</td>
<td>Initiate making friends</td>
<td>1.3</td>
<td>13.8</td>
<td>4.5</td>
<td>14.8</td>
<td>6.5</td>
<td>14.8</td>
</tr>
<tr>
<td>E-23</td>
<td>Get life in a dull party</td>
<td>1.7</td>
<td>9.0</td>
<td>4.0</td>
<td>10.6</td>
<td>4.9</td>
<td>8.7</td>
</tr>
<tr>
<td>E-27</td>
<td>Keep in background (r)</td>
<td>1.1</td>
<td>10.3</td>
<td>3.8</td>
<td>11.4</td>
<td>5.2</td>
<td>10.0</td>
</tr>
<tr>
<td>E-32</td>
<td>Like Mixing</td>
<td>1.5</td>
<td>18.9</td>
<td>6.5</td>
<td>20.8</td>
<td>9.0</td>
<td>20.8</td>
</tr>
<tr>
<td>E-36</td>
<td>Like excitement</td>
<td>1.7</td>
<td>10.4</td>
<td>4.0</td>
<td>11.0</td>
<td>4.9</td>
<td>10.1</td>
</tr>
<tr>
<td>E-41</td>
<td>Be quiet with others (r)</td>
<td>2.2</td>
<td>13.0</td>
<td>4.1</td>
<td>13.8</td>
<td>5.6</td>
<td>12.1</td>
</tr>
<tr>
<td>E-44</td>
<td>Others think lively</td>
<td>3.9</td>
<td>13.6</td>
<td>5.2</td>
<td>14.7</td>
<td>6.4</td>
<td>12.9</td>
</tr>
<tr>
<td>E-48</td>
<td>Get a party going</td>
<td>2.2</td>
<td>11.9</td>
<td>4.9</td>
<td>12.5</td>
<td>5.5</td>
<td>10.2</td>
</tr>
</tbody>
</table>

Table 4. EPQR-S Item endorsement rates by age-gender, total, and selected studies.

<table>
<thead>
<tr>
<th>EPQR-S Item #</th>
<th>Short Text</th>
<th>% Missing</th>
<th>Age 50-62</th>
<th>Age 63-69</th>
<th>Age 70-96</th>
<th>Total</th>
<th>Canberra*</th>
</tr>
</thead>
<tbody>
<tr>
<td>N-01</td>
<td>Your mood changes</td>
<td>0.4</td>
<td>7.9</td>
<td>2.8</td>
<td>8.6</td>
<td>3.2</td>
<td>7.7</td>
</tr>
<tr>
<td>N-05</td>
<td>Feel &quot;just miserable&quot;</td>
<td>1.0</td>
<td>9.2</td>
<td>2.3</td>
<td>9.5</td>
<td>2.6</td>
<td>7.5</td>
</tr>
<tr>
<td>N-09</td>
<td>Be irritable</td>
<td>1.5</td>
<td>2.4</td>
<td>1.8</td>
<td>2.2</td>
<td>1.4</td>
<td>2.3</td>
</tr>
<tr>
<td>N-13</td>
<td>Feelings easily hurt</td>
<td>1.0</td>
<td>14.5</td>
<td>4.2</td>
<td>13.9</td>
<td>4.9</td>
<td>11.6</td>
</tr>
<tr>
<td>N-17</td>
<td>Feel &quot;fed-up&quot;</td>
<td>1.3</td>
<td>8.8</td>
<td>3.0</td>
<td>9.4</td>
<td>3.4</td>
<td>9.1</td>
</tr>
<tr>
<td>N-21</td>
<td>Call self nervous</td>
<td>0.7</td>
<td>7.1</td>
<td>2.2</td>
<td>7.6</td>
<td>2.5</td>
<td>7.2</td>
</tr>
<tr>
<td>N-25</td>
<td>Worry</td>
<td>1.3</td>
<td>14.4</td>
<td>3.6</td>
<td>13.3</td>
<td>4.5</td>
<td>12.6</td>
</tr>
<tr>
<td>N-30</td>
<td>Consider yourself tense</td>
<td>2.4</td>
<td>6.8</td>
<td>2.0</td>
<td>6.7</td>
<td>2.1</td>
<td>6.1</td>
</tr>
<tr>
<td>N-34</td>
<td>Worry after embarrassments</td>
<td>0.6</td>
<td>13.4</td>
<td>4.3</td>
<td>14.4</td>
<td>4.9</td>
<td>14.4</td>
</tr>
<tr>
<td>N-36</td>
<td>Suffer from nerves</td>
<td>0.5</td>
<td>5.2</td>
<td>1.6</td>
<td>5.5</td>
<td>1.5</td>
<td>5.9</td>
</tr>
<tr>
<td>N-42</td>
<td>Feel lonely</td>
<td>1.4</td>
<td>6.2</td>
<td>1.5</td>
<td>7.3</td>
<td>1.8</td>
<td>7.6</td>
</tr>
<tr>
<td>N-46</td>
<td>Are troubled with guilt</td>
<td>1.3</td>
<td>6.4</td>
<td>1.8</td>
<td>7.7</td>
<td>2.3</td>
<td>7.7</td>
</tr>
</tbody>
</table>

| P-02          | Care what people think (r)        | 0.4       | 5.1       | 2.7       | 6.3       | 3.4   | 6.1       | 3.6      | 27       |          |
| P-06          | Debt worries you (r)               | 0.3       | 1.7       | 1.9       | 1.7       | 1.9   | 1.0       | 1.1      | 9        |          |
| P-10          | Take drugs with strange effects   | 0.3       | 0.2       | 0.2       | 0.1       | 0.1   | 0.0       | 0.1      | 1        |          |
| P-14          | Prefer own way over rules         | 0.9       | 4.5       | 2.9       | 4.0       | 2.4   | 4.9       | 3.1      | 22       |          |
| P-18          | Good manners matter (r)           | 0.4       | 0.6       | 0.4       | 0.6       | 0.1   | 0.6       | 0.5      | 3        |          |
| P-22          | Marriage must be abolished        | 0.2       | 0.3       | 0.1       | 0.3       | 0.1   | 0.4       | 0.1      | 1        |          |
| P-26          | Enjoy co-operation (r)            | 0.5       | 0.4       | 0.3       | 0.5       | 0.1   | 0.4       | 0.2      | 2        |          |
| P-28          | Worry about work mistakes (r)     | 1.1       | 1.9       | 1.3       | 1.7       | 1.2   | 1.9       | 1.4      | 9        |          |
| P-31          | Savings & insurance over-rated    | 0.7       | 3.9       | 1.9       | 4.6       | 2.1   | 5.9       | 2.3      | 21       |          |
| P-35          | Try not to be rude (r)            | 0.4       | 0.4       | 0.3       | 0.6       | 0.4   | 0.7       | 0.2      | 3        |          |
| P-39          | Want others afraid of you         | 0.5       | 0.3       | 0.4       | 0.0       | 0.0   | 0.1       | 0.1      | 1        |          |
| P-43          | Follow rules than go own way (r)  | 2.7       | 5.5       | 2.2       | 6.6       | 2.1   | 5.8       | 2.8      | 25       |          |

* Canberra Longitudinal Study of Aging data (Mackinnon, et al., 1995).

a Endorsement significantly different by gender ($\chi^2<.05$, no control for Type I error).

b Endorsement significantly different by age group ($\chi^2<.05$, no control for Type I error).
The average non-response rate for Extroversion items was 1.7%; the non-response rate ranged from 1% (#11, enjoy meeting people) to 3.9% (#44, others think lively). This was the highest non response rate of the 48 EPQR-S items. This item also had the highest non-response rate for the older adults studied by Mackinnon and colleagues (Mackinnon et al., 1995) who reported a rate of 5.9%. It may be that this particular item is more likely to be viewed as inapplicable by older adults.

Similar to Extroversion items, Neuroticism and Psychoticism items also had mean non-response rates lower than the average of 2% previously reported (Mackinnon et al., 1995). The mean non-response rate of Neuroticism items was 1.1%; non-response rates ranged from 0.4% (#1, your mood changes) to 2.4% (#30, consider yourself nervous). The mean non-response rate of Psychoticism items was 0.7%; non-response rates ranged from 0.2% (#22, marriage must be abolished) to 2.7% (#43, follow rules than go own way). While slightly attenuated, these patterns hold even if the eleven respondents who failed to complete the last 12 items are eliminated.

Together, scale and position explain 34.2% (adjusted $R^2$, $F = 7.112, p < .001$) of the non-response variance. After excluding the few subjects who omitted the last 12 items, the non-response rates of the 48 items were analyzed with ANOVA using scale as a fixed factor and position as a covariate. The mean non-response rate for Extroversion items was significantly higher than mean rates for each of the other three scales, and the mean rate for Neuroticism items was significantly higher than the rate for Psychoticism. Scale explained 25.9% of the variance, and when position was added as a covariate, it explained an additional 8.3%. This is in contrast to Mackinnon and colleagues (Mackinnon et al., 1995) who reported no effect of item position on non-response rate. The finding for position is potentially problematic because it constitutes a partial violation of the MAR assumption of the ML methods we used to handle missing data.
Overall, the item non-response rates were small (0.2% to 3.9%), so the effects of scale membership and item position should not distort measurement parameter estimation for the EPQR-S items. Particular attention can be given to items with high non-response rates if invariance across age or sex does not obtain. Further investigation of the motivation for non-response is beyond the scope of this study; however, there are potentially complex relations between scale membership, item position, non-response, age, and sex. This matter deserves further attention.

**Preliminary EFA Analyses**

_Sibling Dependency._ To investigate potential bias due to sibling dependencies, EFA models were estimated using all respondents as well as using a single-sibling subset. The single-sibling subset (N=2604) was selected by taking all singletons (N=1191) and the first sibling entered for each of the remaining 1413 twin pairs, with the exception of mixed-sex twins. In the case of mixed-sex twins, the male sibling was selected so as to offset the lower participation of males. Figure 2 shows the magnitude and pattern of loadings in these models, as well as a series of additional single-factor (i.e., univariate) EFA models.
Figure 2. Magnitude and pattern of loadings in EFA models.
The P scale showed the greatest amount of reactivity to the smaller sample, (see Figure 2). The E, N, and L scales had very stable patterns of loadings whether one or both siblings were used. While the E, N, and L scales appear to be unaffected by any sibling dependency, it is unclear whether the P scale is responding to sibling dependency or simply the reduced size of the dataset. Several P scale items had the most highly skewed distributions.

The inter-factor correlations did appear to be stable regardless of whether all respondents or only single-siblings were used. Figure 3 shows the inter-factor correlations from the two EFA models. The extroversion and neuroticism factors had a small negative correlation, extroversion and psychoticism had a very small positive correlation, and both neuroticism and psychoticism factors has a very small correlation with the lie factor. There were no notable discrepancies.

Even in the single group EFA model with all 4017 respondents, the cross-tabulation of the 2\textsuperscript{nd} and 11\textsuperscript{th} P scale items had an empty cell (EPQR-S #06, "Would being in debt worry you?"; EPQR-S #39, "Would you like other people to be afraid of you?"). In the same four factor EFA model, but using data from only one sibling, there were four empty-cell conditions, involving six of the 12 P scale items. Table 5 shows the cross-tabulation of EPQR-S #06 with EPQR-S #39.

The number of empty-cell conditions in the three-age multi-group model was also investigated. All but one of the twelve P scale items had an empty-cell condition involving one or more of the other P scale items at least one of the three age groups. The sole exception was

<table>
<thead>
<tr>
<th>Table 5. Cross-tabulation of EPQR-S #06 with EPQR-S #39</th>
</tr>
</thead>
<tbody>
<tr>
<td>#39 - Do you want people to be afraid of you?</td>
</tr>
<tr>
<td>#06 - Would being in debt worry you?</td>
</tr>
<tr>
<td>No</td>
</tr>
<tr>
<td>3575 32</td>
</tr>
</tbody>
</table>

![Figure 3. Inter-factor correlations when using all respondents (lower) and only single-siblings (upper).]
EPQR-S #31 (Do you think people spend too much time safeguarding their future with savings and insurances?) In addition, EPQR-S #39 (afraid of you), had an empty cell when tabulated with two N scale items (#13, Are your feelings easily hurt?; #38, Do you suffer from 'nerves'? ) and three E scale items (#07, Are you rather lively?; #15, Can you usually let yourself go and enjoy yourself at a lively party?; #32, Do you like mixing with people?).

Taken together these investigations suggested that we did not have a sufficiently large sample to conduct unbiased multi-group CFA analyses that included the Psychoticism Scale. The distribution of several P scale items was highly skewed. It is likely that these infrequently endorsed items (e.g., Do you want people to be afraid of you?) tap higher levels of psychoticism than is usually present in non-clinical samples. While some of these zero cell counts might be due to sampling and underrepresented the true expected population count, at least some should accurately represent population values of less than 1/4000 ($p = 0.00025$). If the response patterns of our AARP sample generalize, then binominal probability (using Poisson approximation) suggests that sample sizes in excess of 10,000 would be needed to avoid empty-cells and far larger numbers would be needed to achieve cell-counts greater than five. Figure 4 shows the relation of sample size and probability of 5 or more events when the individual event probability is 0.00025.
The manifest P scale score was used as a covariate (in addition to the L scale score) in subsequent invariance models of the E and N scales. The intent was to preserve the correlation extroversion and neuroticism had in the full four-factor EFA model.

**Multigroup CFA Models**

*Baseline CFA Model.* Consistent with personality research recommendations discussed earlier, our baseline model modified the published structure to permit several salient cross-loadings and correlated uniquenesses. This practice has been recommended for CFA models of personality structure (Aluja et al., 2005; Van Prooijen & Van Der Kloot, 2001). Several cross-loadings were permitted based on inspection of cross loadings in our preliminary EFA models, and modification indices in our multi-group CFA models. Modifications were not retained if they did not result a closer match to the pattern of loadings and inter-factor correlations seen the EFA solution.

*Invariance Models.* Four multi-group CFA models, with different patterns or constrained or fixed measurement parameters were estimated across three age groups (50-62, 63-69, 70-96), and also across sex. The least constrained model was the *uniqueness* baseline described previously. The next model added constraints to force loadings to be equivalent across groups. The next model added threshold equivalency constraints. A final model specified across-group

<table>
<thead>
<tr>
<th>Group</th>
<th>Model .. (Constraints)</th>
<th>Free Parm</th>
<th>$X^2$</th>
<th>df</th>
<th>$X^2 / df$</th>
<th>CLI</th>
<th>TLI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>1.. Uniqueness baseline</td>
<td>231</td>
<td>1347.1</td>
<td>378</td>
<td>3.6</td>
<td>.962</td>
<td>.972</td>
<td>.044</td>
</tr>
<tr>
<td></td>
<td>2.. Uniqueness + loadings</td>
<td>183</td>
<td>1248.5</td>
<td>382</td>
<td>3.3</td>
<td>.966</td>
<td>.975</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td>3.. Uniqueness + loadings + thresholds</td>
<td>139</td>
<td>1263.1</td>
<td>396</td>
<td>3.2</td>
<td>.966</td>
<td>.976</td>
<td>.040</td>
</tr>
<tr>
<td></td>
<td>4.. Threshold baseline + loadings</td>
<td>187</td>
<td>1344.0</td>
<td>390</td>
<td>3.4</td>
<td>.962</td>
<td>.973</td>
<td>.043</td>
</tr>
<tr>
<td>Age</td>
<td>1.. Uniqueness baseline</td>
<td>154</td>
<td>1269.7</td>
<td>251</td>
<td>5.1</td>
<td>.959</td>
<td>.970</td>
<td>.045</td>
</tr>
<tr>
<td></td>
<td>2.. Uniqueness + loadings</td>
<td>130</td>
<td>1154.4</td>
<td>252</td>
<td>4.6</td>
<td>.963</td>
<td>.974</td>
<td>.042</td>
</tr>
<tr>
<td></td>
<td>3.. Uniqueness + loadings + thresholds</td>
<td>108</td>
<td>1173.8</td>
<td>259</td>
<td>4.5</td>
<td>.963</td>
<td>.974</td>
<td>.042</td>
</tr>
<tr>
<td></td>
<td>4.. Threshold baseline + loadings</td>
<td>132</td>
<td>1263.4</td>
<td>256</td>
<td>4.9</td>
<td>.959</td>
<td>.971</td>
<td>.044</td>
</tr>
</tbody>
</table>
equivalency for both thresholds and loadings (a loading-restricted version of the threshold baseline, which did not fix uniqueness to 1.0 except as required in the 1st group. Invariance hypotheses involving each measurement parameter were tested by comparing two models that differed only in the constraints on that parameter. Table 6 shows the number of free parameters and fit information for each model.

All models fit the data well. The $\chi^2/df$ was greater than 2.0 for all models, but this is common for large samples and large models with many parameters. In other words, even small amounts of misfit are significant. The usual practice is to rely on relative fit indices (Bentler, 1990; Browne & Cudeck, 1993; Hu & Bentler, 1999). For the CLI and TLI indices, values above .95 indicate a good fit. For the RMSEA, the $a$ value less then .050 is considered to indicate good fit. For all models in both age and sex groups, the CLI ranged between .959 and .066. For the TLI the range was .970 to .975. Finally, for the RMSEA, the range was .040 to .045.

Table 7 summarizes the invariance hypothesis tests using nested full and reduced model comparisons. The $\chi^2$ value for the WLSMV estimator used in this study cannot be used for $\chi^2$ difference tests. An appropriate test (with adjusted degrees of freedom) is given in the MPlus output, and is described in the Mplus Technical Appendices at www.statmodel.com. Change in parameters is included because the $df$ is an adjusted value MPlus provides when using the WLSMV estimator. It does not correspond one-to-one with parameters.

<table>
<thead>
<tr>
<th>Group</th>
<th>Full / Reduced Table 6 #s</th>
<th>Δ free parms</th>
<th>Δ $X^2$</th>
<th>df</th>
<th>$X^2/df$</th>
<th>p-value</th>
<th>Δ Relative Fit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CFI</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>1 / 2</td>
<td>48</td>
<td>71.3</td>
<td>43</td>
<td>1.7</td>
<td>0.004</td>
<td>.004</td>
</tr>
<tr>
<td>Strong</td>
<td>2 / 3</td>
<td>44</td>
<td>52.7</td>
<td>38</td>
<td>1.4</td>
<td>0.056</td>
<td>.000</td>
</tr>
<tr>
<td>Strict</td>
<td>4 / 3</td>
<td>48</td>
<td>69.4</td>
<td>43</td>
<td>1.6</td>
<td>0.007</td>
<td>.004</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak</td>
<td>1 / 2</td>
<td>26</td>
<td>22.2</td>
<td>21</td>
<td>1.1</td>
<td>0.390</td>
<td>.004</td>
</tr>
<tr>
<td>Strong</td>
<td>2 / 3</td>
<td>22</td>
<td>58.9</td>
<td>19</td>
<td>3.1</td>
<td>0.000</td>
<td>.000</td>
</tr>
<tr>
<td>Strict</td>
<td>4 / 3</td>
<td>24</td>
<td>25.9</td>
<td>21</td>
<td>1.2</td>
<td>0.211</td>
<td>.004</td>
</tr>
</tbody>
</table>
All invariance hypotheses were accepted. For all but the test of strong invariance (i.e., loadings and thresholds) across age groups, the $\chi^2/df$ was less than 2.0. However, the change in all three of the relative fit indices was negligible for each test. In fact only small improvements were seen. This is because the constraint introduced very little misfit relative to the degrees of freedom gained when the additional parameters were constrained. Based on simulation work with traditional CFA models (Cheung & Rensvold, 2002), a decrease of 0.10 in CFI has been recommended as the threshold of rejection for invariance hypotheses.

**Factor Mean Trends and Covariances.**

As would be expected with a finding of strict factorial invariance, the factor mean trends and dispersion across age and sex are very similar to the manifest scores. Figure 5 provides a box-plot comparison of raw scale and factor scores for Neuroticism by age group and sex. Figure 6 does the same for Neuroticism. Mean effects across age group for both Extroversion and Neuroticism are small and the distributions have substantial overlap. Males in all three age groups have a lower mean Neuroticism score. For Extroversion, there is no clear sex effect. Older females and younger males have the lowest mean scores. However, here too the distributions substantially overlap the other age groups.

Table 8 shows the estimated correlation of the extroversion and neuroticism factors, the raw scale correlation, and the factor score correlation. This agrees with the -.24 correlation in the EFA model. In all cases, the raw score correlations are lower, and the factor score correlations are higher, which may reflect slight factor score indeterminacy. The correlation of extroversion and neuroticism is stronger for females than males and weaker in the oldest group.

| Table 8. Correlation of Extroversion and Neuroticism across age and gender. |
|-------------------------------|-------------------|-------------------|-----------------|------------------|----------------|
| Source                        | Male  | Female | Young | Middle | Old   |
| Model Estimate                | -0.22 | -0.28  | -0.27 | -0.29  | -0.24 |
| Raw Scale                     | -0.16 | -0.20  | -0.21 | -0.21  | -0.16 |
| Factor Scores                 | -0.25 | -0.31  | -0.30 | -0.33  | -0.26 |
Figure 5. Comparison of raw scale and factor z-scores for Neuroticism
Figure 6. Comparison of raw scale and factor z-scores for Extroversion
Discussion

The analyses here are perhaps the most robust examination of the measurement properties of the EPQR-S in older adults. Mean summary scores, item endorsement rates, and rates of non-response were tabulated by age and sex and compared to values in the published literature. The EPQR-S’ dichotomous item response format has been appropriately modeled with polytomous EFA and CFA models. Multigroup polytomous CFA models were used to test hypotheses of strict factorial invariance across age and sex in older adults. Innovative methodology was employed to permit testing hypotheses of threshold invariance that would not have been tested using the procedure outlined in the literature (Millsap & Yun-Tein, 2004). Factor scores were generated and compared in terms of mean trends, dispersion, and inter-factor correlation with the manifest summary scores and SEM model estimates.

Findings of no, or only small, mean level effects is unexpected given reports of changes in neuroticism and extroversion levels with declines in health and social restrictions that are generally seen with advances age (Maiden et al., 2003). However, health information was not available in our AARP sample. There were considerable individual difference around both the E and N mean levels in both age and sex groupings, and the score distributions were skewed. This may indicate a mixture of two or more distinct populations, perhaps distinguished by health or other aspects of successful aging. Future studies should investigate the invariance of the EPQR-S across levels of physical ability and other conditions of poor health.

Age-Cohort Confound. Interpretation of cross-sectional results of factorial invariance across age groups is, of course, limited by a confound between age and birth cohort. It is also probable that the older individuals are sampled from a more restricted population than the younger individuals, an issue of selective survival (Baltes, 1968). Nevertheless, evaluation of
factorial invariance constitutes an important empirical test of the comparability of measurement that is often assumed in cross-sectional analysis involving means and covariances. Since factorial invariance does obtain across age and sex groups then we can be reasonably confident that the personality factors are comparable in groups comprised of different ages and drawn from different birth cohorts.

Limitations. Several circumstances would strengthen this study. A larger sample would further alleviate concerns about bias due to low frequencies in item cross-tabulations, and permitted the evaluation of the Psychotocism scale. Also, longitudinal data and health information would have permitted a more rigorous examination of changes in personality in older adults. The self-selected nature of our AARP sample may be why higher mean levels on some EPQR-S scales were discrepant from published reports of similarly aged adults in other countries. Further understanding of the meaning and use of the Lie scale is also desirable. Finally, there are several methodological issues that will need further study. For example, when conducting invariance tests in polytomous CFA models, is it justified to use fit-index hypothesis rejection thresholds that were developed for CFA models with continuous indicators? Having said these things, this study should still be taken as a rigorous and methodologically sound investigation of the EPQR-S measurement properties in older adults.
## Appendix I – Full Text of EPQR-S Items

<table>
<thead>
<tr>
<th>Scale/Item</th>
<th>Item Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-01</td>
<td>Are you a talkative person?</td>
</tr>
<tr>
<td>E-02</td>
<td>Are you rather lively?</td>
</tr>
<tr>
<td>E-03</td>
<td>Do you enjoy meeting new people?</td>
</tr>
<tr>
<td>E-04</td>
<td>Can you usually let yourself go and enjoy yourself at a lively party?</td>
</tr>
<tr>
<td>E-05</td>
<td>Do you usually take the initiative in making new friends?</td>
</tr>
<tr>
<td>E-06</td>
<td>Can you easily get some life into a rather dull party?</td>
</tr>
<tr>
<td>E-07</td>
<td>Do you tend to keep in the background on social occasions?</td>
</tr>
<tr>
<td>E-08</td>
<td>Do you like mixing with people?</td>
</tr>
<tr>
<td>E-09</td>
<td>Do you like plenty of bustle and excitement around you?</td>
</tr>
<tr>
<td>E-10</td>
<td>Are you mostly quiet when you are with other people?</td>
</tr>
<tr>
<td>E-11</td>
<td>Do other people think of you as being very lively?</td>
</tr>
<tr>
<td>E-12</td>
<td>Can you get a party going?</td>
</tr>
<tr>
<td>N-01</td>
<td>Does your mood often go up and down?</td>
</tr>
<tr>
<td>N-02</td>
<td>Do you ever feel 'just miserable' for no reason?</td>
</tr>
<tr>
<td>N-03</td>
<td>Are you an irritable person?</td>
</tr>
<tr>
<td>N-04</td>
<td>Are your feelings easily hurt?</td>
</tr>
<tr>
<td>N-05</td>
<td>Do you often feel 'fed-up'?</td>
</tr>
<tr>
<td>N-06</td>
<td>Would you call yourself a nervous person?</td>
</tr>
<tr>
<td>N-07</td>
<td>Are you a worrier?</td>
</tr>
<tr>
<td>N-08</td>
<td>Would you call yourself tense or 'highly-strung'?</td>
</tr>
<tr>
<td>N-09</td>
<td>Do you worry too long after an embarrassing experience?</td>
</tr>
<tr>
<td>N-10</td>
<td>Do you suffer from 'nerves'?</td>
</tr>
<tr>
<td>N-11</td>
<td>Do you often feel lonely?</td>
</tr>
<tr>
<td>N-12</td>
<td>Are you often troubled about feelings of guilt?</td>
</tr>
<tr>
<td>P-01</td>
<td>Do you take much notice of what people think?</td>
</tr>
<tr>
<td>P-02</td>
<td>Would being in debt worry you?</td>
</tr>
<tr>
<td>P-03</td>
<td>Would you take drugs which may have strange or dangerous effects?</td>
</tr>
<tr>
<td>P-04</td>
<td>Do you prefer to go your own way rather than act by the rules?</td>
</tr>
<tr>
<td>P-05</td>
<td>Do good manners and cleanliness matter much to you?</td>
</tr>
<tr>
<td>P-06</td>
<td>Do you think marriage is old-fashioned and should be done away with?</td>
</tr>
<tr>
<td>P-07</td>
<td>Do you enjoy co-operating with others?</td>
</tr>
<tr>
<td>P-08</td>
<td>Does it worry you if you know there are mistakes in your work?</td>
</tr>
<tr>
<td>P-09</td>
<td>Do you think people spend too much time safeguarding their future with savings and insurances?</td>
</tr>
<tr>
<td>P-10</td>
<td>Do you try not to be rude to people?</td>
</tr>
<tr>
<td>P-11</td>
<td>Would you like other people to be afraid of you?</td>
</tr>
<tr>
<td>P-12</td>
<td>Is it better to follow society's rules than go your own way?</td>
</tr>
<tr>
<td>L-01</td>
<td>If you say you will do something, do you always keep your promise?</td>
</tr>
<tr>
<td>L-02</td>
<td>Were you ever greedy by helping yourself to more than your share of anything?</td>
</tr>
<tr>
<td>L-03</td>
<td>Have you ever blamed someone for doing something you knew was really your fault?</td>
</tr>
<tr>
<td>L-04</td>
<td>Are all your habits good and desirable ones?</td>
</tr>
<tr>
<td>L-05</td>
<td>Have you ever taken anything (even a pin or button) that belonged to someone else?</td>
</tr>
<tr>
<td>L-06</td>
<td>Have you ever broken or lost something belonging to someone else?</td>
</tr>
<tr>
<td>L-07</td>
<td>Have you ever said anything bad or nasty about anyone?</td>
</tr>
<tr>
<td>L-08</td>
<td>As a child were you ever cheeky to your parents?</td>
</tr>
<tr>
<td>L-09</td>
<td>Have you ever cheated at a game?</td>
</tr>
<tr>
<td>L-10</td>
<td>Have you ever taken advantage of someone?</td>
</tr>
<tr>
<td>L-11</td>
<td>Do you always practice what you preach?</td>
</tr>
<tr>
<td>L-12</td>
<td>Do you sometimes put off until tomorrow what you ought to do today?</td>
</tr>
</tbody>
</table>
References

Aluja, A., Garcia, O., & Garcia, L. F. (2004). Replicability of the three, four and five
Zuckerman's personality super-factors: Exploratory and confirmatory factor analysis of
the EPQ-RS, ZKPQ and NEO-PI-R. *Personality & Individual Differences, 36*(5),
1093-1108.

factor structure across exploratory and confirmatory factor analyses. *Personality and
Individual Differences, 38*(8), 1879-1889.

organizational research. *Organizational Research Methods, 1*(1), 45-87.

Baltes, P. B. (1968). The logical status of age as an experimental variable: Comments on some
methodological issues. In K. W. Schaie (Ed.), Theory and methods of research on aging
(pp. 172 173). Morgantown, WV: West Virginia University Press.


*Personality & Individual Differences, 5*(6), 615-632.

Personality Questionnaire: An examination of the factorial similarity of P, E, N, and L

107*(2), 238-246.

A study on the 5-factor model of personality. *Personality and Individual Differences,
11*(5), 515-524.


Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor
covariance and mean structures: The issue of partial measurement invariance. 


Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and


Muthen, B. O., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. Psychometrika, (You can get a copy of this paper by emailing bmuthen@ucla.edu. It was never revised and resubmitted).


Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*(2), 139-158.


Chapter 5: Afterword

This dissertation is focused on measurement models and measurement equivalence, and has broadly considered measurement issues that arise when clinical checklists (or other measures with limited response options) are adapted for use in individual differences research. Comparing the task of diagnostic classification as case/non-case to the task of quantifying intra-individual and inter-individual differences, two qualitatively different endeavors are revealed. In more narrow terms, this work has primarily investigated polytomous confirmatory factor analysis (CFA) models, which show promise for the psychometric study checklist instruments with dichotomous yes/no response options, or limited response options (i.e., polytomous). The Developmental Behavior Checklist (DBC, (Dekker, Nunn, & Koot, 2002; Einfeld & Tonge, 1995) examined in Chapter 3, and the Eysenck Personality Questionnaire, short form revised (EPQ-RS, (Eysenck, Eysenck, & Barrett, 1985) studied in Chapter 4 are good examples of instruments developed for clinical classification, and later pressed into use to study intra-individual differences in longitudinal studies. In furtherance of these objectives, considerable attention has been given to factorial invariance (FI) procedures – especially recent extensions to traditional FI procedures that are able to test FI hypotheses for instruments with polytomous items. In Chapter 1 traditional FI procedures are reviewed and issues relating to FI procedures for polytomous models are introduced. In Chapter 4, FI tests are conducted for the Extroversion and Neuroticism subscales of the EPQ-RS; and baseline models, model identification issues, and procedures are described in sufficient detail to guide other researchers in applying these techniques.

While detailed discussion, and where appropriate, detailed procedures, have been provided, many of the psychometric issues raised throughout this work, and ultimately full
understanding of the limitations of FI procedures, require consideration of the nature of the underling constructs and the purpose of the measurement. The factor-analytic model requires that the latent construct be a plausible cause of the observed indicators, and Chapter 1 reviews this issue. Also, the tolerance for changes in patterns of factor loadings, or degrees of acceptable measurement invariance is ultimately a qualitative issue because the latent constructs breadth and level of aggregation must be considered (Bagozzi & Edwards, 1998). This concern is conceptually illustrated in Chapter 1’s discussion of the Satisfaction with Life Scale (SWLS, (Diener, Emmons, Larsen, & Griffin, 1985), and empirically illustrated in Chapter 3’s findings regarding the DBC’s Social Relating and Communication Disturbance subscales. For example, the reversal of the highest and lowest loadings on the Social Relating factor [Underactive (#48), Her own (#57)] in our simple structure CFA model might be regarded as less important since the items both invoke non-social activity. In contrast, the reversal of the highest and lowest loadings on the Communications Disturbance factor [Routine (#5), Stands (#78)] might not be ignorable if the items are judged to have sufficiently distinct meanings in the context of disturbed communications. The Communication Disturbance factor exhibits a definite shift away from the verbal/communication items. If structural equation models containing latent factors are to be used, attention to conceptual shifts between the latent factors and published structures should be considered – especially if the intent is to quantify small differences individuals may exhibit (from themselves) over time.

While models of the relations among constructs in our theories receives a lot of attention, necessary models between our actual data and the constructs they purport to

![Figure 1. Factor-analytic measurement model](image)

Figure 1. Factor-analytic measurement model
measure often receive little attention. A measurement model, like the factor-analytic
measurement model shown in Figure 1, articulates the relation between observed measurements
and the latent constructs they purport to measure. Consistent with the dictum, “all models are
incorrect, some models are useful”, a useful measurement model impose few restrictions,
provides a basis to assess construct validity, and provide a means to assess bias and invariance.
FI testing is one tool to assess bias, but it is incumbent on the researcher to insure that the
constructs are appropriate for a factor-analytic framework, and that any decisions based on
invariance findings consider the breadth of the construct, its level of aggregation, and the
purpose of the measurement. In short, the researcher must consider how incorrect?, how useful?

To complicate matters, there are unanswered questions regarding traditional FI
procedures, and additional questions and unresolved issues particular to polytomous FI
procedures. Measurement Equivalence or Invariance (ME/I) is the broader concept that
subsumes Factorial Invariance (FI). Like the point and the plane, it is a mathematical abstraction.
In reality there are only approximations. It is incumbent on each measurement theory to provide
a methodology for demonstrating ME/I. Factorial Invariance is an ME/I approximation that can
be tested with multigroup confirmatory factor analysis (CFA). (For a review of the literature
see,(Bontempo & Hofer, in press; Vandenberg & Lance, 2000).

Meredeth (1993) defined measurement equivalence as the condition where individuals
with equivalent true scores would have the same probability of a particular observed score on an
associated test. He defined a weaker condition called weak measurement invariance where the
expectation, or on average, such individuals would score the same. It is not possible to
unambiguously demonstrate measurement equivalence in a factor analytic framework, but strict
factorial invariance (SFI), when all measurement parameters function equivalently in each group
of a multigroup CFA model does denote weak measurement invariance for instruments that have continuous and normally distributed item response formats. However, in practice the limited response formats of most instruments used in social science research are weak approximations of continuous and normally distributed. Dichotomous and polytomous response formats offer the most serious violations of this assumption. More research is needed to ascertain that SFI in polytomous models provides the same evidence for weak measurement equivalence as SFI in traditional models.

The invariance models of the EPQ-RS in Chapter 4 highlight some of these unresolved issues. The only discussion in the literature concerning baseline models for FI testing in multigroup CFA models with dichotomous indicators makes recommendations that negate the testing of threshold parameters in the measurement model (Millsap & Yun-Tein, 2004). This is at odds with our reasoning that it is crucial to test the invariance of threshold parameters because they are the sole parameters that connect the measurement model to the actual observed data. Recall that polytomous models (Figure 2) posit an underlying continuous latent response variable for each item. N–1 thresholds, where n is the number of response options, connect the measurement model to the actual data. So, it should be essential that invariance testing include threshold parameters. In Chapter 4, a possible solution to this issue is addressed through the use of two baseline models and two corresponding nested model sequences.

Figure 2. Polytomous factor-analytic model.
Overall, the framework articulated here, the conceptual issues reviewed in Chapter 2 and the empirical illustrations provided in Chapters 3 and 4, should prove useful to researchers who are trying to establish a sound psychometric basis when adapting polytomous instrument originally developed for clinical screening purposes to the more demanding task of quantifying intra-individual differences over time or across other contexts.

References


VITA: Daniel E Bontempo

EDUCATION

- 1981 Bachelor of Science in Computer Science (BS). Loyola University of the South, New Orleans, Louisiana.
- 1992-1993. Thirty hours of undergraduate psychology credit. Texas A&M University, College Station, Texas.
- 1996 Masters of Arts in Counselor Education (MA). Sam Houston State University, Huntsville, Texas.
- 2005 Masters of Applied Statistics (MAS). The Pennsylvania State University, University Park, PA
- 2006 PhD in Human Development. The Pennsylvania State University, University Park, PA

RECENT TEACHING & PROFESSIONAL EXPERIENCE

- 2006- Present. Research Associate. The Penn State Prevention Research Center.
- 1997-2003. Graduate Research Assistant, Graduate Teaching Assistant, Graduate Lecturer. The Pennsylvania State University, College of Health and Human Development.
- 1993-1997 Lab Data Specialist Texas A&M University, College of Architecture, Environmental Psychophysiology Lab. Duties included building lab apparatus, automating lab experiments and data collection, data reduction and analysis.

PUBLICATIONS


