# Complementary and competing factor analytic approaches for the investigation of measurement invariance

ANDREA HILDEBRANDT, OLIVER WILHELM and ALEXANDER ROBITZSCH

Sample-related invariance is an important topic in psychometric research. The generalizability of findings in a broad range of application samples requires equivalence of interpretations based on the measurement outcomes across respective samples. Contextual factors like gender, age, culture, ethnicity, socio-economical status etc. may affect the meaning and interpretation of psychological measures. Sample-related invariance is frequently investigated using Multiple-Group Mean and Covariance Structure (MGMCS) analyses. This method builds upon natural or artifical categories of contextual variables. Many contextual variables are continuous variables and their categorization is associated with an information loss and potentially overly simplistic data analyses. We present and discuss two complementary analytical approaches – Latent Moderated Structural (LMS) Equations and Local Structural Equation Models (LSEM). Both approaches allow treating contextual factors as continuous variables and are appropriate to detect non-linear relations. The use of these methods is exemplified based on real data. We investigated measurement equivalence of a battery of cognitive tests across age ($N$ = 448; age range 18-82 years). Based on a higher-order factor model of cognitive abilities factorial equivalence could be established – contradicting the age-dedifferentiation hypothesis. Advantages and disadvantages of MGMCS, LMS, and LSEM and further implementations beyond aging-research are discussed.

*Key words*: measurement invariance, factorial invariance, lifespan, validity, assessment

"*The scientist is usually looking for invariance whether he knows it or not. Whenever he discovers a functional relationship his next question follows naturally: under what conditions does it hold*?" (Stevens, 1951)

As pointed out in the quotation by Stevens (1951), the issue of invariance is a highly relevant topic in scientific measurement. Concepts of invariance and the development of methods to determine whether or not specific aspects of equivalence are hold for a given measurement instrument have a long history in psychometrics and can be traced back to Thorndike (1904) and Thurstone (1925, 1947; see Engelhard, 1992 for a historical review). There is a bifocal perspective on the concept of invariance in psychometrics, that Thurstone emphasized and that was later immersed by Rasch (1961). One perspective is that of *sample invariance* of a measure and the other perspective is that of *item invariant measurement* of individuals (see also Engelhard, 1992). Evidence of sample invariance of a measure or a single test item is given, when subgroup characteristics (like gender, age, social class etc.) do not influence ability estimates or an item scale value. Item invariant measurement refers to the concept of minimizing discrepancies between ability estimates of an individual when using for example two different item pools. In the present paper we will focus on the issue of sample-related invariance.

The main goal of this paper is to present and illustrate two analytical approaches for the investigation of measurement invariance. To this end we will first shortly review the concept of cross-group (sample-related) psychometric invariance defined in factor analytical context and name some of the prevailing research questions linked to its investigation. Second will be the description of the commonly used method to study invariance (Multiple-Group Mean and Covariance Structures – MGMCS) and the discussion

Andrea Hildebrandt, Department of Education, Humboldt University at Berlin, Unter den Linden 6, D-10099 Berlin, Germany.
E-mail: andrea.hildebrandt@iqb.hu-berlin.de (the address for correspondence and requests of program codes).

Oliver Wilhelm, Department of Education, Humboldt University at Berlin, Germany;

Alexander Robitzsch, Federal Institute for Education Research, Innovation & Development of the Austrian Schooling System (BIFIE Salzburg).

*Acknowledgments*

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (Wi 2667/2-3) to Oliver Wilhelm and Werner Sommer.

of some disadvantages of this method particularly in the context of investigating developmental research questions. The two complementary usable analytical approaches will then be presented. These methods allow a more stringent investigation of specific aspects of invariance whenever group or sub-sample characteristics are continuous variables. This outline will be followed by an implementation of the two approaches for cognitive ability measures in an age heterogeneous sample. We will conclude by discussing advantages and disadvantages of the proposed methods and by considering further possible implementations beyond aging research.

### *Measurement and factorial invariance*

In Confirmatory Factor Analysis (CFA) the question of invariance is traditionally the one of construct comparability across discrete groups (sub-samples at different developmental levels, like young vs. older individuals, males vs. females, or groups of different culture, ethnicity etc.). However, for example in the case of age the "grouping variable" is indeed a continuous variable, whenever observations on a broader age range are available. Cases like culture or ethnicity might in fact be a similar case (Betancourt & Lopez, 1993). Usually such naturally continuous variables are categorized for analytical purposes in the literature. But categorization brings along loss of information and therefore has, other things being equal, less power to detect true differences (MacCallum, Zhang, Preacher, & Rucker, 2002; Preacher, Rucker, MacCallum, Nicewander, 2005). Obviously, the categorization of continuous variables is also associated with poor sensitivity to detect onsets of changes or other discontinuities.

### *The importance of testing sample-invariance at the level of observed variables*

There are plenty of practical reasons to ask whether a test is measuring the same construct, independent of discrete or continuous contextual factors characteristic for sub-samples investigated with it. Ethnicity, culture, social status, gender and age are perhaps the most frequently investigated factors which potentially affect measurement outcomes. In order to establish fair and unbiased assessments of constructs the goal of test development is to reduce the influence of such factors regarding the meaning of the measures. Evidence for barred influences can be provided by establishing *measurement invariance*, which is a precondition of cross-group comparisons on latent factor level (Meredith, 1993). Measurement invariance testifies that "reliable measurement properties have been defined in the same operational manner" (Little, Card, Slegers, & Ledford, 2007, p. 125) independently from contextual "grouping" factors.

But what does measurement invariance exactly mean? Approximately one and a half decades after the influential work by Meredith (1993), the practical and theoretical guide on studying measurement invariance by Horn and McArdle (1992) and the hot and informative discussions (see Labouvie & Ruetsch, 1995b; McDonald, 1995; Meredith, 1995; Nesselroade, 1995a, 1995b) on this topic, elicited by the work of Labouvie and Ruetsch (1995a), who suggested a relaxation of previously established demands of invariant factor loadings as prerequisite of measurement invariance, it was Little and coworkers who elaborated a scientifically updated and very useful guide on measurement and factorial invariance (see Little et al., 2007). The definitions we will introduce and the terminology we will use in the present paper, mainly rely on the work by Little et al. (2007).

There are three levels of invariance implied by the term of measurement invariance: 1) *Configural invariance* – is substantiated if a given set of observed variables are shown to be indicators of the same construct(s) across disjoint samples. This means that the number of factors extracted from a set of observed variables is the same across sub-samples. But configural invariance also implies an equivalent pattern of factor loadings. For example, if an indicator requires dual-loadings in a specific sample, whereas only one factor accounts for the systematic variance of the respective indicator in another sample-group, the assumption of configural invariance is affected; 2) *weak invariance* (also known as *metric invariance*) – requires equally sized factor loadings across groups and entails therewith a first level of quantitative invariance. But note that factor loadings are just relatively equal at the weak invariance level, because factor variances can still vary across groups. Hence factor loadings are weighted by the difference in latent variances at this level. A further step of establishing measurement invariance in the factor analytic approach is the 3) *strong invariance* (also called *scale invariance*) – which is demonstrated when loadings and intercepts of the indicators are equivalent across groups. Note that equivalence of residual variance of the indicators – referred to in the literature as 4) *strict invariance* – is not needed for measurement invariance, because the residuals do not contain reliable, construct common variance of the indicators (see Little et al., 2007; Meredith, 1993).

### *The importance of testing sample-invariance at the level of latent constructs*

Once measurement invariance (demonstrated by strong invariance and not strict) was established cross-group comparisons on the level of constructs can be meaningfully carried out. Theoretical considerations can raise further factorial invariance questions beyond those of the measurement, now at the level of constructs (latent factors). There are at least three latent level parameters which are usually of interest in theoretically guided cross-group comparisons or comparisons as a function of continuous contextual variables: 1) factor variances, 2) between factor covariances (correlation), and 3) factor means.

In order to illustrate age-group comparisons of factor variances and covariances as they are frequently found in the literature, we will consider the influential *differentiation-dedifferentiation* hypothesis, elaborately investigated in the developmental research of cognitive abilities (Lindenberger & Baltes, 1997; Reinert, 1970; Tucker-Drob & Salthouse, 2008). The hypothesis postulates changes of the organization of intelligence across the lifespan that go beyond mean improvements during childhood and youth and mean decline in old age: Performances on a set of cognitive tasks in early childhood can be described by a narrower factor space, which differentiates during childhood development. This means that the number of factors explaining cognitive performances increase until early adulthood. Whereas adult age is characterized by stability of the factorial structure of cognition, a reintegration or dedifferentiation follows during old age. We will further consider only the dedifferentiation part of the hypothesis for our exemplification. A strict form of dedifferentiation postulate that the number of cognitive factors will decrease with aging. This would imply that not even configural invariance could be established in older groups compared to groups of younger adults. However, cross-sectional and longitudinal lifespan data do not support strict factorial dedifferentiation (Cunningham, 1981; Schaie, Willis, Jay, & Chipuer, 1989; Schaie, Maitland, Willis, & Intrieri, 1998). Testing a weak form of dedifferentiation transfers the hypothesis to the domain of parameters of the latent level, because the question is whether covariances (correlations) between cognitive factors and also their variance increase in old age. Furthermore, invariance related hypotheses regarding latent level covariances are also highly relevant in any validation context.

There are plenty of studies on mean level comparisons across groups in many psychological research fields (gender difference, cross-cultural, developmental studies etc.). Unfortunately most of these studies are based on single tasks or small samples or both. Consequently, conclusions on the level of abilities – as opposed to specific tasks – can not be drawn on their basis. An up-to-date method of investigating cross-group mean differences on the level of constructs is to test such differences after factorial invariance was established for the measures. Factorial invariance means that the constructs can be equally interpreted in all groups. Hence, testing mean level differences after establishing measurement invariance precludes that "apples and oranges will be compared".

### *Analytical approaches of testing age-invariance*

The approaches of investigating measurement and factorial invariance reviewed further below will be exemplified on a cross-sectional lifespan sample. For this reason, we will elaborate our discussion of the methods especially for the case of testing age-invariance and point out further possible applications later on. The used techniques of invariance testing are all implemented in the context of Structural Equation Modeling (SEM) or CFA. For details on SEM modeling see for instance Kline (2005).

*Multiple-Group Approaches.* The traditional procedure to investigate measurement and factorial invariance is the modeling of Multiple-Group Mean and Covariance Structures (MGMCS; Little et al., 2007). In MGMCS analysis a defined structure is fitted across different groups in a series of models with varying restrictions of parameter-equality with the general goal to determine whether the structure is the same or not across groups. This simultaneous modeling allows testing whether specific parameters (factor loadings; intercepts; latent factor correlations) can be restricted to have the same value in the subgroups. Testing measurement invariance consequently requires the comparison of a series of nested models (see Bollen, 1989). In a first step (*configural invariance*) the same model is fitted but all model parameters are allowed to vary across groups. In a second step factor loadings are constrained to be equal across groups (*weak invariance*) – the metric invariance model is a nested version of the configural model. In a third step (*strong invariance*) a nested model of the weak invariant model is tested, which constrains intercepts to be equal across groups. Based on the $\chi^2$-difference test the amount of loss of fit as a consequence of parameter restrictions can be tested (Bollen, 1989). Due to the fact, that $\chi^2$-values are highly sensitive to large sample sizes and number of constraints, further alternative indices of evaluating deterioration due to restrictions were developed. For example, Browne and Du Toit (1992) proposed a rescaling of the $\Delta\chi^2$-value in the metric of RMSEA. The authors called the rescaled value *Index of Root Deterioration per Restriction* (RDR). MGMCS models in which restrictions lead to values higher than RDR=.08 should not be considered invariant.

MGMCS approaches are methodologically sophisticated tools and should be promoted to conduct cross-group comparisons on measures in the case of discrete (categorical) grouping variables. However, there are some disadvantages of MGMCS models for the study of age-related invariance or changes of individual differences across the lifespan. Age is obviously a continuous variable, but in MGMCS approaches it is treated as a categorical variable. Usually a relatively high number of age points are compressed into one value (15-20 years or even more), and so within-group observations can represent developmentally strongly different states. Furthermore, the category boundaries defining the groups are frequently arbitrarily defined. One could only overcome the problem of compression by collecting a high number of observations within a narrowly defined age range (e.g. 5 years). Such cross-sectional design would easily imply very large samples (e.g. 1,000 participants or more) if a larger age range is investigated (e.g. age range 20-80 years). This is rarely feasible, particularly in expensive laboratory studies.

In the following, we want to illustrate several models using the equation of the dependency of an indicator *Y* on a latent factor *F* modeled in the measurement model part of the SEM:

$$Y= \lambda_0 + \lambda_a F + \varepsilon \qquad , a=1, ..., A \tag{1}$$

where $\lambda_a$ denotes the factor loading of *Y* on the latent factor *F* in one of the *A* age groups *a*. Note, that if metric invariance does not hold, the factor loadings are allowed to differ in different age groups and are implicitly modeled as a regression step function which realizes different values for the different age groups, but a constant factor loading within all ages of one age group.

To overcome these disadvantages of MGMCS analyses a less frequently used (but see Tucker-Drob, 2009, for an exception) but very useful analytical approach – Latent Moderated Structural Equations (Klein & Moosbrugger, 2000) – can be applied in the context of invariance testing. With this method the equivalence of specific parameters from CFA or SEM models can be estimated avoiding the rather arbitrary discretization of the age variable. Another fruitful procedure to investigate the onset and shape of age-related changes in factorial structures or factor levels with relatively low sample sizes at any age point is the modeling of a series of Age-Weighted Measurement or Local Structural Equation Models using weighting functions. We will further describe these two methods below and apply them to real data in the empirical section of this paper.

*Latent Moderated Structural Equations.* A recently developed analytical approach – Latent Moderated Structural Equation (LMS; Klein & Moosbrugger, 2000) – implemented in the statistical software Mplus (Muthén & Muthén, 1998-2007), can be complementary used to MGMCS analyses to test metric and factorial age-related invariance. Usually, in CFA or SEM the indicators (observed variables) are linearly related to the latent factors. In age heterogeneous samples it is however possible that the regression slopes (loadings of the indicators) are moderated by age. Thus their magnitude might change as a function of age. In terms of MGMCS this would be visible as a lack of age-related metric invariance. Furthermore, as already mentioned above, SEM approaches are used to test linear relationships (correlation or regression) between latent variables. But these relationships can be moderated by age or nonlinear transformations of age. This would be visible as evidence in support of differentiation or dedifferentiation of ability space as discussed above. Using LMS one can test, whether the influence of an exogeneous (independent) latent variable on an endogeneous (dependent) latent variable is moderated by age or transformations of age and thus investigate factorial invariance in a given model.

Such moderation effects can be tested by defining an interaction term between the exogeneous latent variable and a continuous observed variable (e.g. age) and regress the dependent variables onto the interaction term. Dependent variables are indicators in the case of testing metric and endogeneous latent variables if the aim is to investigate factorial invariance. Formally, this can be written in the following two equations:

$$Y= \lambda_0 + \lambda_1 (\mathrm{a})\, F+ \varepsilon \tag{2}$$

$$\lambda_1 (\mathrm{a}) = \beta_0 + \beta_1 a + \beta_2 a^2 \tag{3}$$

Equation 2 assumes a linear relationship of the indicator on the latent factor, whereas in equation 3, *F* is regressed on age using additionally a quadratic regression. Plugging equation 3 into equation 2, this leads to the combined equation

$$Y= \lambda_0 + \lambda_1 \beta_0 F+ \lambda_1 \beta_1 Fa + \lambda_1 \beta_2 Fa^2 + \varepsilon \tag{4}$$

The interaction terms of factor *F* and age *a* indicate the dependence of the factor loading on age. The reasercher has to determine if the regression coefficients $\beta_1$ and $\beta_2$ are reasonably high to reject the hypothesis of a constant factor loading across age.

Similar to the case of MGMCS models where fit indices of more or less restricted models are compared with each other in order to prove whether the imposed restrictions deteriorate the fit, in the case of LMS, models with and without interaction effects are compared. If the exclusion of the interaction impairs the fit of the model this constitutes an indication for lack of metric or factor covariance invariance.

LMS models are limited compared to linear SEM models in that many established fit statistics (like CFI, RMSEA etc.) are not applicable to them. In regular SEM models the $H_0$ is a restricted and the $H_1$ an unrestricted covariance matrix between which the discrepancy is estimated. However, interaction models generate non-linear outcomes where sample covariance matrices are not sufficient statistics any more. In the absence of popular fit indices the interaction vs. non-interaction model can be compared using likelihood ratio tests or penalty functions, both based on the $\chi^2$-statistic.

Apart from their limitation concerning the lack of fit indices, LMS models have a substantial advantage over MGMCS models in testing invariance across the lifespan because they allow treating age as a continuous variable. LMS models dispel any kind of disadvantages related to the problem of categorization of continuous variables. Taking another point of view, LMS "smoothes out" the regression step functions in multiple-group models and can be seen as parametrizations of structural variations in MGMCS. In principle, also more complex nonlinear functions of age than in equation 3 can be specified if sample size permits this.

*Age-Weighted Measurement or Local Structural Equation Models.* A further analytical approach we wish to introduce in this work is the use of Age-Weighted Measurement or Local Structural Equation Models (LSEM). Such models can be very informative in testing invariance questions, because they allow the description of age-gradients of parameter estimates from CFA and SEM models across a wide age-range based on relatively small age heterogeneous samples. Locally-weighted averaging, used in nonparametric regression (Fox, 2008) and nonparametric mixed effects models (Wu & Zhang, 2006), can be implemented in SEM context in order to define observation weights as a function of age and fit a series of models using differently weighted

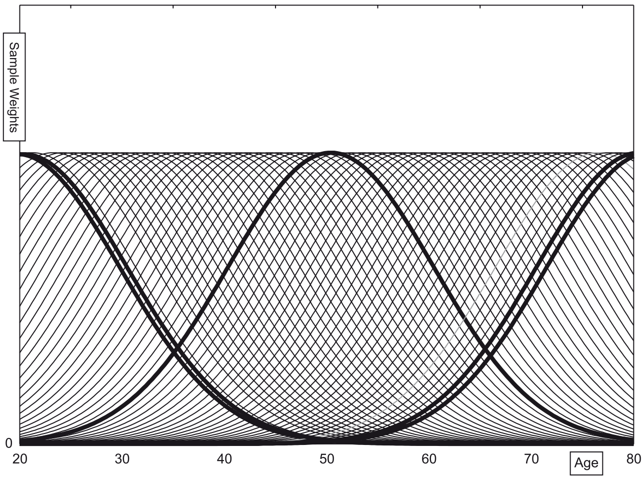

*Figure 1.* Course of defined sample weights – schematic illustration.

observations. This approach also allows treating age as a continuous variable.

In the empirical exemplification we will use a kernel function of weighting observations (Gasser, Gervini, & Molinari, 2004) around focal (central) age points. Note that alternative weighting functions could also be implemented (Wu & Zhang, 2006). Focal age points can be defined even in one year steps. Let us assume that age based sample weights for a focal point of 50 years shall be calculated which is illustrated by the solid curve in Figure 1. Observations at the focal point of 50 years receive the highest weight. Using the kernel weighting function sample weight of observations around a focal age point can be calculated. Calculated weights will fall off symmetrically with increasing distance of an observation from the focal value which has the maximum weight in the case of using a normal kernel function. The idea is that ages nearby the focal point of 50 years are used to borrow information for the calculation of the SEM on that focal point and ages far distant from 50 years should have negligible influence. Using this calculated sample weights, a SEM with weighted observations is being estimated.

Based on calculated sample weights for a series of focal age points (see calculation steps of weights in the results section of this paper), CFA or SEM models can be sequentially fitted moving the weighting window along the age variable. Parameter estimates and fit indices for the series of models can further be plotted against age in order to visualize their equivalence or change. Age related changes can than be computed for the parameter estimates and the fit indices. In essence, pursuing our illustration example, the factor loadings are allowed to smoothly vary for all different ages $a$:

$$Y = \lambda(a)F + \varepsilon \quad (5)$$

Equation 5 highlights the difference to the MGMCS and the LMS approach. In MGMCS the loading function was discretely defined and is a (nonparametric) regression step function (see equation 2). In LMS models a parametric loading function is being imposed (see equation 4). Therefore, both MGMCS and LMS can be regarded as approximations of the local SEM. The deviation of the loading function λ($a$) from a constant function (say, the calculation of a standard deviation λ($a$)) can be interpreted as an effect size of specific parametric non-invariance.

### *Empirical illustration – higher-order-structure of cognitive abilities*

In this section we will exemplify the introduced analytical approaches of testing measurement and factorial invariance on the basis of a higher-order structure model of cognitive abilities.

## METHOD

### *Sample*

The data for this illustration are performance measures from a sample of 448 individuals with a mean age of 49 years (*SD* = 20) and a heterogeneous educational background. Half of the participants were female. The sample can be divided into three almost equally sized sub-samples of young (Age Range = 18-35 years, $M_{age}$ = 24, *SD* = 5), middle-aged (Age Range = 36-64 years, $M_{age}$ = 49, *SD* = 8), and older adults (Age Range = 65-82 years, $M_{age}$ = 72, *SD* = 5). Gender distribution does not vary across age groups, all consisting of 50% females. Educational background of the participants was heterogeneous in each age group and comparable to each other, except the slightly more positively selected older group in the direction of a somewhat higher proportion of participants with academic degrees. To evaluate general cognitive functioning of the older participants the Mini-Mental State Examination (MMSE; Folstein, Folstein, & McHugh, 1975) was administrated. No individual performed below the cut-off score of 24 taken to indicate increased risk of mild cognitive impairment (Folstein et al., 1975; Small, Viitanen, & Bäckman, 1997; Schramm et al., 2002).

### *Procedure*

The data were collected within a large cross-sectional aging study on face cognition abilities (Hildebrandt, Sommer, Herzmann, & Wilhelm, submitted). The data relevant for this paper are ability measures on abstract cognitive tests. Each task was separately instructed directly before its administration. Tasks were programmed using Inquisit 2.0© and conducted on PCs with 17 inch color monitors, 85 Hz refreshing rate, 1280 × 1024 resolution, and a viewing distance of approximately 50 cm.

*Tasks*

*Working Memory*

*Memory updating* (MU). This task of measuring working memory was adapted from Oberauer, Süß, Schulze, Wilhelm, and Wittmann (2000). In a three by three grid matrix – depending on the memory load of a trial – two to seven cells were white, whereas the others were black. One-digit numbers consecutively appeared for one second in each white cell. Participants memorized the numbers and their localization in the grid. Following the presentation of the last digit, four arrows pointing vertically up or down appeared in the white cells, one at a time. If upward-pointing arrows appeared in a cell participants mentally added one to the digit a priory presented in that cell and maintain the new number. Downward-pointing arrows instructed participants to decrease the digit by one and to maintain the new number. After the last computing instruction participants had to recall the updated numbers for each white cell. In the recall phase question marks appeared one at a time and the final number for the specific cell had than to be typed in. Participants worked on five practice and 18 experimental trials. During practice, feedback was provided.

*Rotation Span* (RS). A second working memory task was adapted from Shah and Miyake (1996). The task required memorizing and recalling a sequence of arrows, while concurrently engaging in a secondary task of letter-rotation. Each item consisted of a sequence of alternated storage and processing trials. First, an arrow was presented, which radiated out from the centre of a circle and showed in one of eight possible directions (up, down, left, right, diagonal left up, diagonal left down, diagonal right up and diagonal right down). In addition arrows were short or long, so that 16 possible arrows resulted. Presented arrows had to be memorized. Following the presentation of an arrow a normally or mirror-reversed letter (“G”, “F”, or “R”), rotated at 0, 45, 90, 135, 180, 225, 270, or 315 degrees was displayed. Participants indicated if presented letters were normally displayed or mirror-reversed. After response a second arrow was presented, which also had to be memorized. The arrow was again followed by a letter-decision trial. Memory load (list length) of the items varied between two and five arrow-letter pairs. At the end of a sequence of letter-arrow pairs, a graphic depicting the 16 possible arrows appeared. Participants used the mouse to indicate the arrows they memorized by clicking on the corresponding points of the answer screen. Recall was required in the correct order of presentation. A total of 12 items were presented.

*Raven’s Advanced Progressive Matrices* (RAV)

Sixteen items from the original full test (Raven, Court, & Raven, 1979) were included. Items consisted each of a three by three matrix of symbols with the bottom right hand symbol missing. Participants had to choose the symbol that logically completed the matrix from eight options presented below the matrix.

Participants older than 65 years worked only on 11 of the item sequence administered for young and middle aged persons. Elderly worked on five easier items from *Raven’s Standard Progressive Matrices*. Scoring in this task is consequently based on a linked 2-Parameter Logistic Model (see *Scoring and data treatment* section of the paper for details).

*Immediate and Delayed Memory*

Three tasks, each with immediate and delayed recall were our measures of memory. One of these tasks, which required memorizing of word pairs and recalling a second word when confronted with the first one, was taken from the Wechsler Memory Scale (WMS; Härting et al., 2000). The task was slightly modified and computerized. The number of trials was increased from six to eight in order to circumvent possible ceiling effects, because the original six trials were expected to be too easy for mentally healthy young participants. In this work, we will refer to the immediate recall part of this task as *Verbal Memory Immediate* (VMI) and to the delayed part as *Verbal Memory Delayed* (VMD).

A second memory task was adapted based on the WMS, and required memorizing first and last names and recalling the surname when the first name was presented (*Name Memory Immediate* - NMI). Eight pairs of first and last names were used. There was also a delayed recall after approximately one and a half hours (*Name Memory Delayed* – NMD).

Finally, a third memory task was adapted using the same procedure applied in the WMS. In this task pairs of street names and house numbers were learned and immediately recalled after learning (*Address Memory Immediate* – AMI) and again recalled after a delay time of approximately one and a half hours (*Address Memory Delayed* – AMD).

*Mental Speed*

*Finding A’s* (FAs). German words were presented in this task, one at a time. Participants had to decide whether the displayed word contains an “A” or not and responded as quickly as possible by pressing a labeled key on the left if words did not contain an “A” and a key on the right hand side if they contained one. We administered six practice trials with accuracy feedback followed by 80 test trials.

*Symbol Substitution* (SyS). One of the following four symbols appeared in the middle of the screen: “?”, “+”, “%”, or “$”. Participants were required to respond by pressing the upward-pointing arrow key to “?”, the right-pointing arrow key to “+”, the down-pointing arrow key to “%”and the left-pointing arrow key to “$”. There were six practice trials with feedback on accuracy, and 80 test trials.

*Number Comparison* (NC). Two number strings, varying from 3 to 13 digits in length, were presented in each trial. Participants were required to decide whether or not the number strings were identical or differed in one number and to press the corresponding button. There were 6 practice trials with feedback about the accuracy of the decision, followed by 80 test trials.

### *Scoring and data treatment*

Performance indicators in working memory tasks were defined as the average of remembered stimuli at the correct position across all items. Memory tasks were scored as the average of correct responses in the immediate recall and delayed recall respectively.

Scores on the Raven's progressive matrices included in the structural analyses are ability estimates from a 2-Parameter Logistic Model (2PL), since 31% of the completed items were different for participants older than 65 years. We replaced the five most difficult items administered to the younger participants with easy items taken from the standard matrices, in order to avoid floor effects and frustration in the older group. In a 2PL-Model, person (ability) parameters are estimated using logistic *item characteristic curves*, which connect observed responses to continuous latent traits. The estimation of a person parameter is based on the response pattern of the individual, taking the item difficulty and the item discrimination parameter into account (Schmiedek, 2005).

Parameters of interest for the mental speed indicators were the averages of the inverted latencies obtained across all correct responses, calculated as 1000 / reaction time in milliseconds. The scores can be interpreted as number of correctly processed trials per second. To minimize the influence of outliers, before calculating the inverted latencies, reaction times smaller than 200 ms were set to missing. Response latencies 3.5 intraindividual *SDs* above the individual mean were fixed to the individual mean value plus 3.5 intraindividual *SDs*. This procedure was repeated as long as there was no latency with a value above the individual mean plus 3.5 intraindividual *SDs* left. In no case more than 20% of the intraindividual reaction times had to be replaced by their mean plus 3.5 intraindividual *SD*s.

## RESULTS

### *The Higher-Order Factor Model of Cognitive Abilities*

There are many different CFA models introduced in the intelligence literature aimed to represent the structure of individual performance differences in cognitive ability tests (Schulze, 2005). One frequently used structure – particularly in the area of age related changes in cognitive abilities (Tucker-Drob, 2009) – is the higher-order factor model (see Figure 2). This model postulates a hierarchical organization: There are several narrow first-order ability factors, directly linked to the indicators (observed performance in the tasks), and a single second-order ability factor capturing the communality of the first-order factors. The second-order factor is usually labeled as general cognitive ability (g) or fluid intelligence ($g_f$). Higher-order factors are considered to have greater nomological breadth than first-order factors, because they are linked indirectly to more observed variables. The proportion of unexplained variance in a first-order factor is called residual or disturbance and they are represented by a d and the corresponding factor label in Figure 2.

In order to explain performances in the 12 tasks administered in the present study, we postulated three first-order factors and one second-order fluid ability factor. Latent factors are represented as ellipses in Figure 2, observed variables as rectangles, and factor loadings are depicted as directed arrows. First-order factors are: *Immediate and delayed memory* (IDM) – modeled by the three memory tasks, each having an immediate and a delayed recall condition (six indicators; with expected correlations between error terms of two indicators from one task); *mental speed* (MS) – based on three speed tasks, each with one condition of administration (three indicators) and *reasoning/working memory* (REA) – based on the Raven's matrices score and two tasks of working memory (three indicators). The rationale behind using working memory tasks as indicators of reasoning ability relies on strong evidence in the literature of very high latent level correlations between working memory and fluid intelligence (reasoning) – see the seminal work of Kyllonen and Christal (1990) or the reanalysis of the meta-analytic data from Ackerman, Beier, and Boyle (2005) conducted by Oberauer, Schulze, Wilhelm, and Süß (2005). The authors report a latent level correlation of $r = .85$. Schmiedek, Hildebrandt, Lövdén, Wilhelm, and Lindenberger (2009) also found very strong correlations in a recent study based on a single sample (.78-.84, depending on the working memory tasks used to model the working memory factor).

Before proceeding to the modeling of the postulated higher-order model of cognitive abilities across age-groups we fitted the model for the young sub-sample, with participants' age ranging between 18-35 years. This model serves as a baseline and tests whether the postulated structure is true for the present young sample. This analytical step is important because most studies that established the higher-order structure of cognitive abilities rely on young samples. Misfit in the young sample would compromise the intended invariance analyses. We scaled the latent factors in this baseline model by fixing their variance to one.

The model for the young sample fitted the data very well: $\chi^2$ [48] = 60.5, $p = .10$, ($N = 149$), CFI = .98, RMSEA = .04, SRMR = .06. Standardized factor loadings of the reasoning / working memory indicators ranged between $\lambda = .67 – .73$, of the memory indicators between $\lambda = .50 – .86$, and of the speed indicators between $\lambda = .57 – .88$. All were significant at $p < .01$. Loadings of the first-order factors on the higher

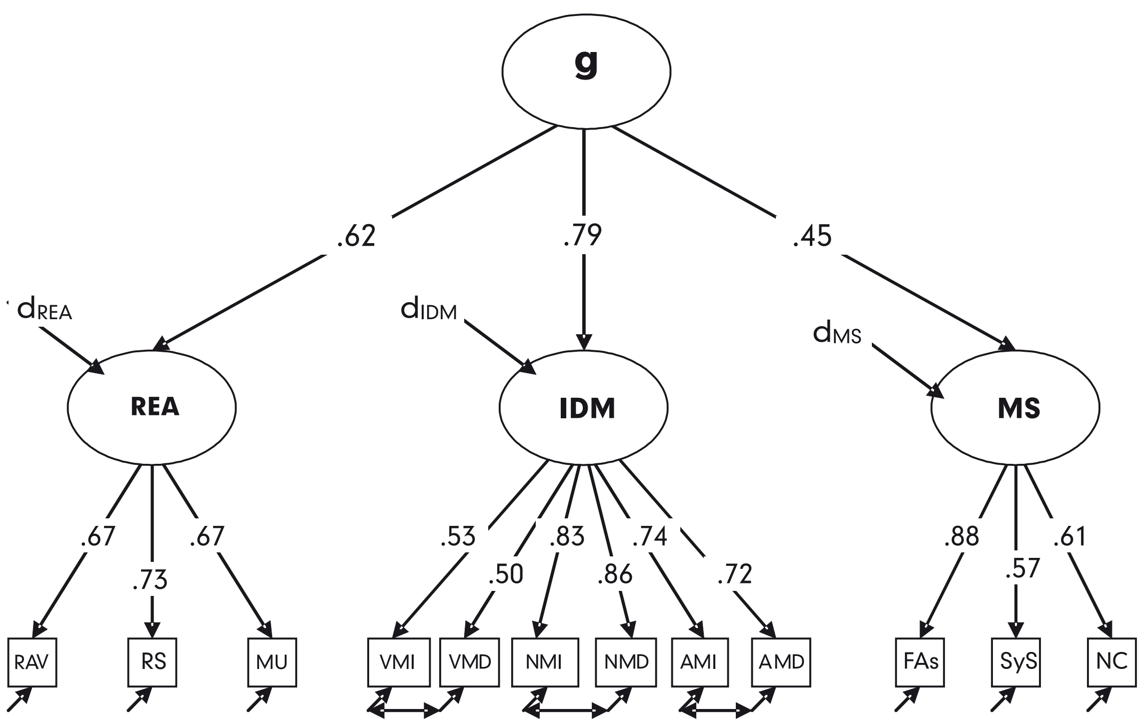

*Figure 2.* Higher-order structure model of cognitive abilities

*Note*. g – General Cognition; REA – Reasoning; IDM – Immediate and Delayed Memory; MS – Mental Speed; R – Residual variance, not accounted for by g; *RAV* – Raven; *RS* – Rotation Span; *MU* – Memory Updating; *VMI* – Verbal Memory Immediate; *VMD* – Verbal Memory Delayed; *NMI* – Name Memory Immediate; *NMD* – Name Memory Delayed; *AMI* – Address Memory Immediate; *AMD* – Address Memory Delayed; *FAs* – Finding A's; *SyS* – Symbol Substitution; *NC* – Number Comparison.

order factor of *general cognition* are $\lambda = .62$ ($R^2 = .39$) in the case of reasoning, $\lambda = .79$ ($R^2 = .63$) in the case of memory and $\lambda = .45$ ($R^2 = .21$) for mental speed. Given the good fit of this model in the young sample this structural representation – depicted in Figure 2 – will be used in the following sections to test invariance and exemplify the analytical approaches discussed in this paper.

### *Invariance Testing using Multiple-Group Analyses*

As discussed above, the traditional method of testing measurement invariance in the factor analytic approach is the estimation and evaluation of MGMCS models. In a first step, the baseline model established in the young sample (see Figure 2) was fitted for three age groups of young, middle-aged and older adults. In this model we postulated the same factorial and loading pattern depicted in Figure 2 across the three age-groups (configural invariance). We scaled latent factors using a recently developed method by Little, Slegers, and Card (2007). Based on this method latent scales are identified by fixing the loadings for each latent variable to an average of one and the sum of indicator intercepts to zero. This has the advantage that all indicator loadings, factor variances, and factor means are freely estimated in all groups. The configural model fitted the data well: $\chi^2$ [144] = 212.5, $p < .01$, CFI = .969, RMSEA = .06, SRMR = .06, supporting the first step of measurement invariance defined by an equal number of factors and the same pattern of factor loadings across groups. In this model the parameters are allowed to vary in their magnitude across sub-samples. Table 1 displays freely estimated standardized factor loadings of the indicators on the first-order factors across age-groups. By inspecting Table 1, strong variations of loadings can be noticed for two of the speed indicators (FAs and SyS). The loadings of the remaining indicators show small variation across age groups. In the metric invariant model we will test if the observable disparities in the magnitude of the loadings estimated in the configural model are statistically significant or not.

For this purpose in a second MGMCS model unstandardized factor loadings of all indicators were constrained to be equal across groups (metric invariance). This model still fitted the data reasonably well: $\chi^2$ [162] = 256.2, $p < .01$, CFI = .958, RMSEA = .06, SRMR = .09, but the $\Delta\chi^2$ of 43.7 corresponding to a difference of 18 degrees of freedom between the configural and the metric model reached significance.

*Table 1*
Standardized factor loadings of the indicators on the first-order factors (configural model - MGMCS)

| Task | Reasoning | | | Task | I & D Memory | | | Task | Mental speed | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Young | Middle | Older | | Young | Middle | Older | | Young | Middle | Older |
| RAV | .67 (.06) | .59 (.06) | .70 (.05) | VMI | .53 (.06) | .52 (.07) | .63 (.06) | FAs | .88 (.07) | .64 (.06) | .74 (.06) |
| RS | .73 (.06) | .75 (.05) | .65 (.06) | VMD | .50 (.06) | .54 (.07) | .64 (.06) | SyS | .57 (.06) | .81 (.07) | .61 (.05) |
| MU | .67 (.06) | .74 (.05) | .76 (.05) | NMI | .83 (.05) | .71 (.06) | .70 (.05) | NC | .61 (.06) | .60 (.07) | .75 (.05) |
| | | | | NMD | .86 (.04) | .74 (.06) | .75 (.04) | | | | |
| | | | | AMI | .74 (.05) | .76 (.06) | .78 (.04) | | | | |
| | | | | AMD | .72 (.05) | .63 (.07) | .76 (.05) | | | | |

*Note. RAV* – Raven; *RS* – Rotation Span; *MU* – Memory Updating; *VMI* – Verbal Memory Immediate; *VMD* – Verbal Memory Delayed; *NMI* – Name Memory Immediate; *NMD* – Name Memory Delayed; *AMI* – Address Memory Immediate; *AMD* – Address Memory Delayed; *FAs* – Finding As; *SyS* – Symbol Substitution; *NC* – Number Comparison; *SE*s are presented in brackets.

Thus the strict statistical test does not support metric invariance across age. However, as pointed out above, the $\Delta\chi^2$-values are highly sensitive if a large number of constraints is estimated in larger samples. For such cases Browne and Du Toit (1992) recommended a rescaling of the $\Delta\chi^2$-values into the RMSEA metric. This Index of Root Deterioration per Restriction (RDR, calculated as √ $\Delta\chi^2$- Δdf / Δdf * *N*) can be interpreted like an RMSEA coefficient. Values below .05 indicate that the difference in fit can be considered of only minor importance. Based on the RDR and also the minor deterioration in other descriptive fit indices (CFI and RMSEA) we prefer the metric invariant over the configural model. We proceed by comparing the weak invariant model with a model testing measurement invariance, by fixing the indicator intercepts to be equal across groups and test, whether strong (scale) invariance is supported by the data. Model fit indices for the stepwise MGMCS models for testing invariance are summarized in Table 2.

Constraining indicator intercepts to be equal across age-groups (strong or scale invariance) considerably affected the model fit: $\chi^2$ [180] = 344.3, *p* < .01, CFI = .927, RMSEA = .08, SRMR = .12. The $\chi^2$-difference test ($\Delta\chi^2$ = 88.1 corresponding to Δdf = 18), the RDR Index (RDR = .09) and the strong loss of fit also expressed by the CFI and RMSEA values all suggest that strong invariance is not supported by the data (see also Table 2).

*Table 2*
Invariance testing across age (MGMCS)

| Model | $\chi^2$ | *df* | CFI | RMSEA | $\Delta\chi^2$ | Δ*df* | RDR |
|---|---|---|---|---|---|---|---|
| Configural invariance | 212.5 | 144 | .969 | .06 | --- | --- | --- |
| Weak (metric) invariance | 256.2 | 162 | .958 | .06 | 43.7** | 18 | .05 |
| Strong (scale) invariance | 344.3 | 180 | .927 | .08 | 88.1** | 18 | .09 |
| Factorial invariance – comparison with the weak invariant model | 262.9 | 166 | .957 | .06 | 6.7 | 4 | --- |

Consequently, testing the invariance of the loadings of the three first-order factors on the higher-order factor was carried out based on the weak invariant model. This model tests whether the common variance of the factors is equal in magnitude across age-groups and is a form of testing the factorial invariance on the first-order level. The age-related invariance or change of these loadings is tested by the dedifferentiation hypothesis (see Tucker-Drob, 2009). A weak form of age-related dedifferentiation would be supported, if loadings were significantly higher in the older compared to younger groups. Fixing the respective loadings to be equal in the factorial invariant model ($\chi^2$ [166] = 262.9, *p* < .01, CFI = .957, RMSEA = .06, SRMR = .09) compared to the weak invariant model, did not significantly affect the goodness of fit, indicated by the non-significant $\Delta\chi^2$-value of 6.7 corresponding to four degrees of freedom difference between the models.

Summing up, MGMCS models show that the configuration of the model depicted in Figure 2 is true for all three age-groups and all non-standardized factor loadings (both λs of the indicators and λs of the first-order factors) can be fixed to be equal across groups (showing weak and first-order level factorial invariance). However, as displayed in Table 1 there are at least slight variations in the estimated λ-parameters in the MGMCS models. As we argued in the introduction, MGMCS models are not viable to investigate the onset and shape of age-related changes of SEM model parameters, at least not with usual sample sizes of less than 200 observations for an age-range of 15-25 years or even more. In order to more adequately illustrate lifespan changes of such parameters by treating age as a continuous variable Age-Weighted Measurement or Structural Models

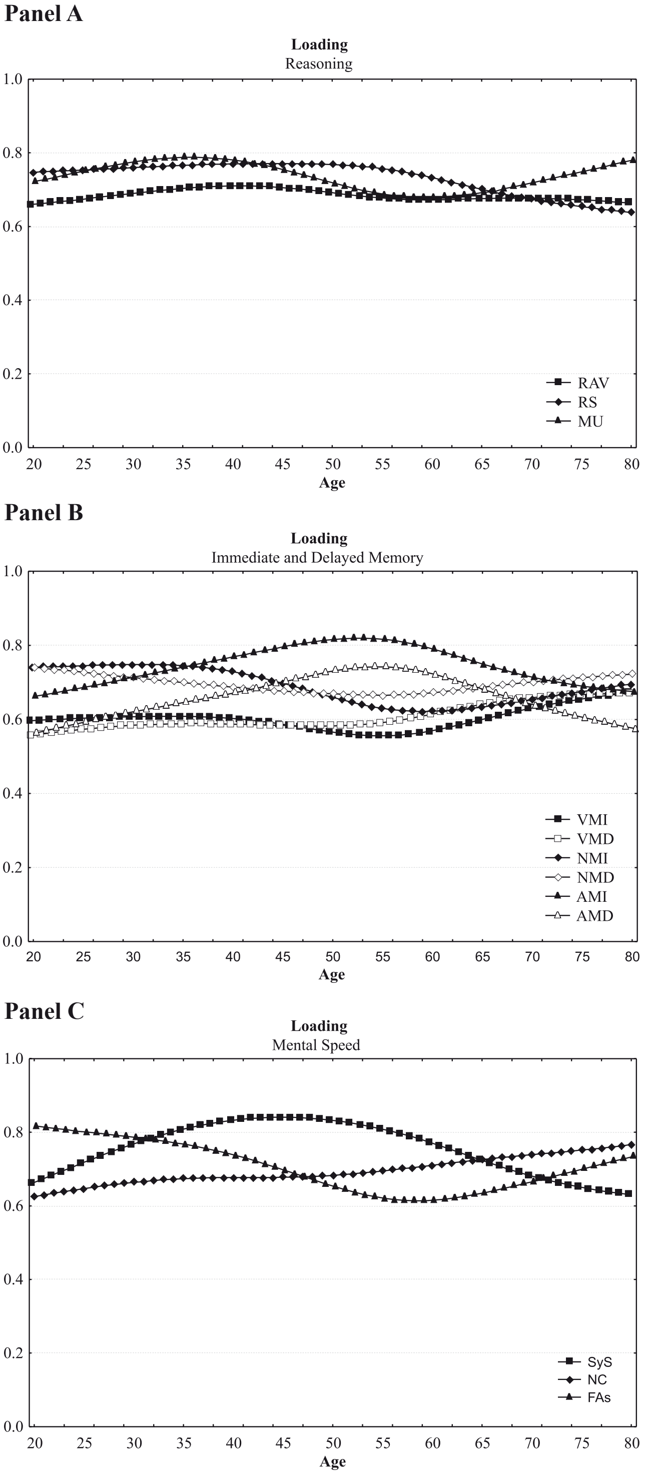

*Figure 4.* Age-gradients of indicator factor loadings – Age-Weighted Models; Panel A: *RAV* – Raven; *RS* – Rotation Span; *MU* – Memory Updating; Panel B: *VMI* – Verbal Memory Immediate; *VMD* – Verbal Memory Delayed; *NMI* – Name Memory Immediate; *NMD* – Name Memory Delayed; *AMI* – Address Memory Immediate; *AMD* – Address Memory Delayed; Panel C: *FAs* – Finding A's; *SyS* – Symbol Substitution; *NC* – Number Comparison.

(Local SEM) can be computed to visualize and describe age-gradients of the estimated parameters.

### *Age-Gradients of Model Parameter form Age-Weighted Models (Local SEM)*

In order to estimate a series of age-weighted models of the baseline model depicted in Figure 2 we computed observation weights around focal age points defined in one year steps from 20 to 80 years. We used a kernel function of weighting observations (Gasser et al., 2004). The following computations were carried out to define sample weights for 61 focal age points between 20 and 80 years:

The bandwidth (bw) of the kernel function was calculated based on the formula:

$$\text{bw} = 2 * N^{(-1/5)} * SD_{\text{age}} \qquad (1b)$$

A scaled distance ($z_x$) was computed by subtracting the focal age from every observation for each focal point:

$$z_x = (\text{age}_x - \text{focal age}) / \text{bw} \qquad (2b)$$

Weights were than calculated based on the normal kernel function for every focal point:

$$K_{\text{focal age}} = (1 / \sqrt{2\pi}) * \exp(- z_x^2/2) \qquad (3b)$$

Finally, weights (W) were rescaled to obtain values between 0 and 1:

$$W_{\text{focal age}} = K_{\text{focal age}} / .399 \qquad (4b)$$

Subsequently – based on sample weights calculated with the formulas 1b-4b – we estimated the higher-order structure model of cognitive abilities with the moving weighting window along the 61 samples of weighted observations.

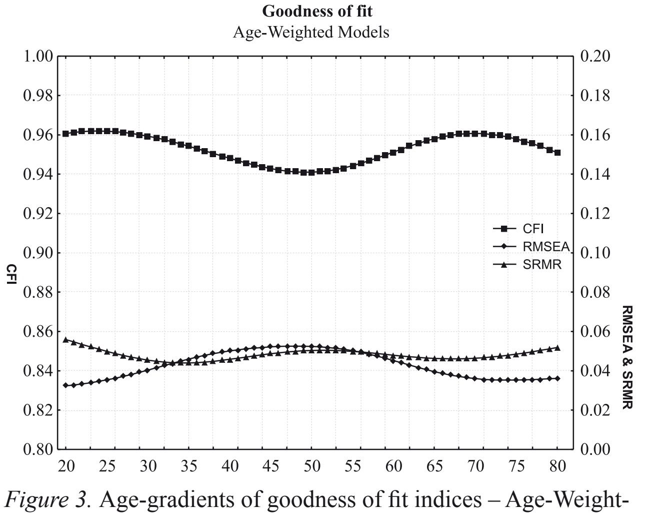

*Figure 3.* Age-gradients of goodness of fit indices – Age-Weighted Models.

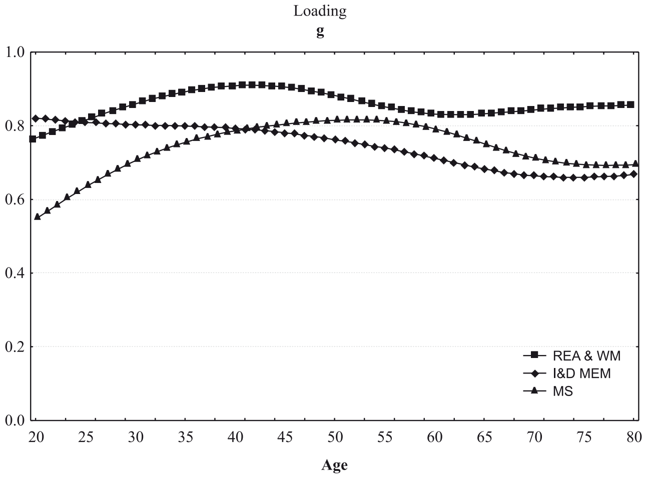

*Figure 5.* Age-gradients of first-order factor loadings – Age-Weighted Models

*Note*. g – General Cognition; REA – Reasoning; WM – Working Memory; IDM – Immediate and Delayed Memory; MS – Mental Speed.

The effective *N* for single computations depended on the frequency of observations at and around the respective focal point. Given varying *N* for the series of models, their goodness of fit will be estimated based on fit indices not affected by sample size: CFI, RMSEA, and SRMR. Figure 3 displays age-gradients of the fit indices estimated for the series of 61 age-weighted models. The fit of the models is particularly good at the beginning and at the end of the age distribution. From 40-60 years the goodness of fit slightly decreased, but the fit indices still indicated reasonable fit.

As we argued above, age-gradients of parameter estimates from such age-weighted models can be very useful in case of lack of invariance, because they provide evidence about possible sources of misfit and are also indicative for the onset of relevant changes. Panel A, B, and C of Figure 4 displays age-gradients of the loadings of the indicators on the first-order factors and Figure 5 loadings of first-order factors on the higher-order factor. Taken together, the gradients show relative stability of factor loadings particularly in the case of reasoning / working memory (Figure 4, Panel A) and delayed memory (Figure 4, Panel B). The loadings of the immediate memory and two mental speed tasks (SyS and FAs, also suggested in MGMCS models; Figure 4, Panel C) show more notable variations across age. Whereas the loadings of the factors for reasoning / working memory and immediate and delayed memory on the general cognition factor (Figure 5) show no relevant age-changes (and therefore do not support the dedifferentiation hypothesis), the loading of mental speed considerable increases up to the age of 60 and slightly decrease thereafter. Note that a so called boundary bias can be responsible for changes at the tails of the age distribution, where the weighting curves become less symmetric. The LMS models we will discuss next might not be affected by this issue. These models will also allow more traditional inferential tests about the significance of the age related changes we report.

### *Testing Metric and Factorial Invariance using Latent Moderated Structural (LMS) Equations*

Metric (indicator loadings) and factorial invariance (first-order factor loadings) were tested in two independent analytic steps using LMS. In a first step, we estimated a correlation model between the first-order factors. The higher-order factor was not modeled directly in this case. Note that the two models (correlation and higher-order structure model) are statistically equivalent. Therefore, their fit is the same. In order to investigate age-related metric invariance using LMS equations, an interaction term between the latent factors and age has to be created. In the higher-order model, the general factor accounts for the common variance of the first-order factors. Creating the interaction between age and a first-order factor in such a model, the interaction will be defined based on the residual of the respective first-order factor (its variance accounted for by the higher-order factor being partialed out). However, the correlation model permits the specification of interactions between age and unresidualized first-order factors, so that the moderation effect with respect to the indicator loadings will be investigated more stringently in this case.

In a second analytic step – aimed to test whether loadings of the first-order factors on the second-order factor change as a function of age – the higher-order model was fitted using LMS. In this case the interaction term was created between age and the second-order factor. The first-order factors were then regressed onto the interaction. This model tested whether loadings between the first-order factors and the second-order factor change as a function of age.

Further, to estimate the influence of quadratic age-effects on the loadings, in both modeling steps we used a non-linear term of the age variable (age-squared) on top of the age variable. For the LMS modeling – in order to reduce issues of multicollinearity – age was centered by subtracting the sample mean from the value of each observation.

We will first describe the results of the first modeling step (the correlation model to test metric invariance). A non-interaction model assumes that indicator loadings do not change across age. Task performances were predicted by linear and quadratic age trends and their linear loadings on the first-order factors in this case. The loglikelihood-value of this model is displayed in Table 3. All linear loadings and linear age trend were significant at $p < .01$. Quadratic age effects were suggested only for reasoning at $p < .01$ ($t = 3.082$) and there was a trend in the case of mental speed ($t = 1.974$, $p = .048$). In the next model we defined interaction terms between all first-order factors and age. Task

*Table 3.*
Loglikelihood-values and comparison between the age and age$^2$-modification model of indicator loadings and the metric invariant model

| | log-likelihood (L) | scaling correction factor (scf) | free parameters (fp) | $\Delta\chi^2$ | $\Delta df$ | AIC | BIC |
|---|---|---|---|---|---|---|---|
| non-modification (A) | 495.62 | 1.05 | 48 | --- | --- | -895.2 | -698.2 |
| age-modification (B) | 536.65 | 1.07 | 60 | 71.36 | 12 | -953.3 | -707.0 |
| age$^2$-modification (C) | 539.72 | 1.05 | 66 | 7.22 | 6 | -947.5 | -676.5 |

*Note.* $\Delta\chi^2 = 2 * (L_B - L_A) / c$; where $c = (scf_B * fp_B - scf_A * fp_A) / (fp_B - fp_A)$

*Table 4.*
Age- and age$^2$-modification effects of indicator loadings expressed as *t*-values

| | RAV | RS | MU | VMI | VMD | NMI | NMD | AMI | AMD | FAs | SyS | NC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| age-mod | .07 | -.30 | **3.31** | 1.31 | **3.72** | -1.63 | -1.17 | .52 | **2.77** | **2.90** | **-2.60** | -.29 |
| age$^2$-mod | --- | --- | 1.42<br>-.18 | --- | --- | -.91<br>.55 | --- | -.53<br>.70 | .35<br>.29 | 1.31<br>-.31 | -.86<br>.09 | --- |

*Note*. Significant effects are printed bold; the second value in the second line of the table displayed the changed parameter estimate for the age-modification term, after the age$^2$-modification term was also added to the model; age-mod – Model with the age modification term of indicator loadings; age$^2$-mod – Model with the added quadratic age modification term of indicator loadings; *RAV* – Raven; *RS* – Rotation span; *MU* – Memory updating; *VMI* – Verbal memory immediate; *VMD* – Verbal memory delayed; *NMI* – Name memory immediate; *NMD* – Name memory delayed; *AMI* – Address memory immediate; *AMD* – Address memory delayed; *FA's* – Finding As; *SyS* – Symbol substitution; *NC* – Number comparison.

performances were than predicted by linear and quadratic age trends, linear loadings and age-modified loadings (interaction term). Table 3 displays the loglikelihood-value of this interaction model. In order to estimate whether there was a significant improvement in model fit in the interaction model we computed the $\Delta\chi^2$-value based on the loglikelihood-values and the scaling correction factors estimated for both models. The difference value ($\Delta\chi^2 = 71.36$, $\Delta df = 12$) is statistically significant. Therefore, including the interaction terms as predictors of task performances did improve the model fit and factor loadings are not age invariant.

A closer inspection of the interaction effects shows which of the 12 loadings change as a function of age. Parameters for the moderation effects are displayed in Table 4. Five of the 12 parameter estimates reach the conventional significance level. Loadings of *Memory Updating, Verbal Memory Delayed, Address Memory Delayed* and *Finding A's* increase and the loading of *Symbol Substitution* decreases with advancing age. However, the age-gradients of indicator loadings estimated on the basis of parameter estimates from the age-weighted models displayed in Figure 4 suggest that there may be some quadratic moderation effect of age. This may be the case for *Memory Updating, Name Memory Immediate, Address Memory Immediate* and *Delayed, Symbol Substitution* and *Finding A's*. In order to test this assumption, in the next model further interaction terms between age-squared and the three latent first-order factors were created. These terms were added as predictors of the performance in the named tasks. Loglikelihood values of this model are displayed in the third line of Table 3. The difference value ($\Delta\chi^2 = 7.22$; $\Delta df = 6$) is not statistically significant and suggests no improvement in model fit. Therefore, adding the age$^2$-interaction terms to the model is not indicated. As shown in Table 4, none of the age$^2$-moderation terms of indicator loadings were significant. However, it should be noted that there was a strong loss in the age-interaction effects in the case of indicator loadings to which also quadratic effects were added. Furthermore, the age-interaction effects were not larger or "more" significant in this case, as displayed by the second value in the second line of Table 4.

A further goal of the present paper was to investigate age-related changes of the loadings of first-order factors in a higher-order model of cognitive abilities as an exemplification of testing factorial invariance using LMS equations. For this purpose, in the first step the hierarchical model was fitted with fixed indicator loadings. These loadings were fixed to values estimated freely in the preceding estimation, in order to reduce minor non-equivalence influences on the indicator level, when testing invariance at the factorial level. The likelihood-value and associated statistics of this model are shown in Table 5. In the second step we defined an interaction term of age and the second-order factor of general cognitive ability and tested whether the interaction modifies the loadings of first-order factors. The loglikelihood of this

*Table 5.*
Loglikelihood-values and comparison between the age and age$^2$-modification model of first-order factor loadings on the higher-order factor and the factorial invariant model

| | log-likelihood (L) | scaling correction factor (scf) | free parameters (fp) | $\Delta\chi^2$ | $\Delta df$ | AIC | BIC |
|---|---|---|---|---|---|---|---|
| non-modification | 495.59 | 1.06 | 39 | --- | --- | -913.1 | -753.1 |
| age-modification | 497.22 | 1.06 | 42 | 3.07 | 3 | -910.4 | -738.0 |
| age$^2$-modification | 498.84 | 1.06 | 45 | 3.06 | 3 | -907.7 | -722.9 |

model is shown in the second line of Table 5. The difference value ($\Delta\chi^2 = 3.07$, $\Delta df = 3$) between the non-interaction and the interaction model is not statistically significant and therefore suggests stability of the loadings across age. We also tested quadratic age-interaction effects, which were statistically not significant as well (see line three of Table 5; an LMS syntax for Mplus was published by Tucker-Drob, 2009).

## DISCUSSION

The goal of this paper was to introduce and illustrate complementary and competing approaches for the investigation of factorial invariance. More specifically, we investigated two analytic approaches that can serve as alternative or complementary approaches to MGMCS analyses, whenever the influence of continuous contextual variables on the meaning of the measurement has to be estimated. Issues of measurement invariance are usually investigated by means of multiple-group analyses. However, problems associated with dichotomizations or categorizations of continuous variables are well established in the methodological literature (MacCallum et al., 2002; Preacher et al., 2005). Therefore, the use of MGMCS is suboptimal in many instances and methodological alternatives need to be developed in order to test for the invariance of constructs for continuous contextual factors. Consequently, we attempted to present and develop useful tools for enhancing research on measurement invariance along naturally continuous contextual factors. We will now review the methods presented here and discuss possible implementation beyond aging research.

### *Assessment of the different methods and possible implementations beyond aging research*

### *Multiple-Group Mean and Covariance Structure Analyses*

Compared to the prevailing practice to investigate group differences by means of ANOVA, which disregard individual differences, MGMCS analyses have the advantage to take within-group individual differences in consideration. Furthermore, abstracting from indicator specificities by estimating latent factors in MGMCS models is very useful too. Nevertheless, these models come along with some disadvantages, we wish to shortly summarize here.

First, in the case of naturally continuous contextual factors – like age, abilities, different trait factors, but also culture, socio-economical status, income, education etc. – the building of category boundaries is highly arbitrary in most cases and categories are in fact artificially created. We want to go one step further and argue that many MGMCS applications are overly simplistic. For example, testing for invariance across ethnical or cultural groups is usually based on self-reported group membership. The problem of what ethnicity really is (Betancourt & López, 1993) on a psychological level and – more important for the present purposes – how ethnicity might affect the construct under investigation is usually neglected. Obviously, it is not the case that the use of continuous variables by itself is a panacea to the problem. Indeed, our own treatment of the meaning of contextual variables is pivotal. What does "age", for example, represent – regardless of whether it is used as categorical or continuous variable? Issues of the meaning of contextual variables and the nature and origin of the relation with the construct under investigation – including potentially theoretically derived reasons for a lack of invariance – are rarely investigated. Is the variable "age" we are interested in really "time passed since birth"? What is the mechanism of the "passing time" and how does it affect "age"? We are convinced however that the use of continuous contextual factors allows a more profound and elaborate theorizing on mechanisms of relations than categorical contextual factors.

Second, in MGMCS-designs it is common that substantial ranges from the continuum of the contextual variables are categorized into one value. Therefore, the method is not well suited to detect onsets and changes along the continuum and it is also unprepared for the parametrization and estimation of nonlinear and interaction effects.

To put the choice of the cut scores for the categorization of a continuous variable in a different light, now from a methodological point of view, we emphasize that a specific categorization *defines* a "contrast" between subgroups. If these prespecified subgroups are compared with respect to invariance, then MGMCS is well suited for this design. We claim that this is rarely the case. If no subgroups are defined a priori, we argue that a sequence of models with varying cut scores for the groups should be estimated to investigate the sensitivity of the invariance measures. If conclusions are stable under various conditions, then no further investigation is needed. However, such a procedure relies on relatively large samples required across the whole frequency distribution. MGMCS compared to LMS have the advantage to conceptualize all aspects of invariance (configural, weak, strong, factorial).

### *Local Structural Equation Models*

In order to work around these two problems we introduced the method of LSEM. The rational behind this method is that observations near a focal value of a contextual factor are more informative for that value than a more distal observation. This is intuitively plausible for continuous variables. The farer away neighboring points are, the less weight they have for that focal point. Age is conceptually opaque. Other continuous variables changing with age might be more plausible contextual factors and of course more sophisticated concepts might be used to investigate invariance. The invariance of intelligence might be more intensively investigated as a function of continuous factors, like ability levels for example. Evidently, the presented methods allow new perspectives on many research questions, including for example the ability differentiation-dedifferentiation hypothesis traced back to Spearman (1927) and recently investigated with a latent interaction approach by Tucker-Drob (2009).

Locally-weighted models have the great advantage of allowing relatively small sample sizes in the investigation of invariance for continuous variables. LSEM also allow the detection and visualization of critical results. They can be used to visualize gradients of loadings, intercepts and latent level parameters along continuous contextual variables and thus test weak (metric), strong (scale) and factorial invariance. Disadvantages of LSEM are the less symmetric weighting functions at the tails of the distribution and the lack of standard inferential tests for global model comparisons. Nevertheless, the amount of deviations from factor loadings and correlations with for example age can be used as an effect size of non-invariance for specific parameters.

To derive inferential tests in LSEM, a researcher could conduct a permutation test approach (Good, 2005). Then, all ages in the sample are randomly paired with persons. In this pseudosample, all relations of loadings, variances and correlations should be independent of age. Conducting this permutation approach, a large number of times (say, 1000 times) and applying LSEM to this pseudosample data leads to a distribution of LSEM parameters under the null hypothesis of no relationship with age. Then, the LSEM age curves can be inspected and pointwise (i.e. for each age) or global tests can be applied.

### *Latent Moderated Structural Equations*

LMS models do allow inferential comparisons not readily available in LSEM. Such models – just as the LSEM models – also allow modeling of continuous variables in an attempt to investigate their influence on the structure of measurement outcomes. But LMS models go beyond locally-weighted structures by additionally allowing traditional inferential statistics based on the loglikelihood of competing models. It will be interesting to see which further model evaluation tools and fit indices for LMS will be developed and recommended in the future.

The disadvantage of categorization disappears in LMS models. However, their usage assumes a specific relationship (linear, quadratic, cubic) of loadings (for example) in order to obtain estimates of regression coefficients of the relationship with age. A linear relationship of a factor loading with age estimates two model parameters effectively, whereas in a MGMCS with three age groups three parameters are estimated such that one more parameter is used to describe the relationship with age. Therefore, LMS can be seen as restricted MGMCS if parametrization are imposed on regressions of model parameters on age groups. At the present, LMS can only be used to test weak (metric) and factorial invariance. If mean structure analysis are included into LMS in the future, they will also applicable to test strong (scale) invariance.

### *Two-Step Procedures: Deriving parameter estimates from LSEM outputs*

The output of a LSEM model can be used to calculate regression parameters or means when nonparametric regressions of factor loadings are regressed linearly or categorized on age. For these derived parameter estimates, a distribution of pseudosamples obtained by the permutation approach defines the distribution under the null hypothesis of no relationship with age. In many cases, an extract of one or some parameters of a nonparametric regression is much easier for communication purposes and possesses a lower standard error than a loading estimate at one specific age. Therefore, an exploratory oriented LSEM can be a starting point for confirmatory analyses.

## CONCLUSIONS

Conclusions derived from our analyses in the empirical section of this paper using the three approaches are converging. Thus the same major interpretations can be drawn on their basis: There are minor age-related variations at the level of indicator loadings and no variation in the loadings of first-order factors. These conclusions were supported by the results from all three analytical approaches. Locally-weighted models allow more detailed descriptions of misfit on the indicator level than MGMCS analyses do. Furthermore, LMS equations facilitate inferential statistical estimations of possible parameter changes and also allow more sophisticated modeling of non-linear trends.

### *Theoretical implications*

Although the analyses were not aimed to test theoretical assumptions, but to exemplify modeling methods for the testing of a specific class of hypotheses (measurement invariance), there is one theoretical implication of our findings on a substantive level we wish to emphasize. As pointed out in the introduction, the age differentiation-dedifferentiation hypothesis has a very long history in the developmental research and its investigation led to controversial discussions. Recently, Tucker-Drob (2009) presented comprehensive analyses of the age and ability differentiation-dedifferentiation hypothesis, by applying LMS for the first time in the literature in order to test these hypotheses. The author did not find support for age-related dedifferentiation (expressed in higher loadings of first-order factors on g with increasing age). The analysis we presented here replicate the findings of Tucker-Drob (2009). LMS and LSEM might be more adequate methods for the investigation of the dedifferentiation hypothesis than the methods used so far, which do not go beyond MGMCS and frequently fall short of a methodologically adequate test of the dedifferentiation hypothesis.

### *Our recommendation*

Closing up, we recommend the use of LMS and LSEM to test invariance in factor-analyses along continuous contextual variables. Extending traditionally used MGMCS analyses with the proposed LSEM and LMS equations, invariance questions can be investigated in a more fine-grained setting and in this way sources of possible lack of invariance or nonlinearity can be estimated more precisely.

## REFERENCES

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin, 131,* 30-60.

Betancourt, H., & Lopez, S. R. (1993). The study of culture, ethnicity, and race in American psychology. *American Psychologist, 48,* 629-637.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.

Browne, M. W., & Du Toit, S. H. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research, 27*, 269–300.

Cunningham, W. R. (1981). Ability factor structure differences in adulthood and old age. *Multivariate Behavioral Research, 16,* 3–22.

Engelhard, G. (1992). Historical views of invariance: Evidence from the measurement theories of Thorndike, Thurstone, and Rasch. *Educational and Psychological Measurement, 52,* 275-291.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-Mental State. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198.

Fox, J. (2008). *Nonparametric simple regressions*. Sage University Papers.

Gasser, T., Gervini, D., & Molinari, L. (2004). Kernel estimation, shape-invariant modeling and structural analysis. In R. Hauspie, N. Cameron, & L. Molinari (Eds.), *Methods in human growth research* (pp. 179–204). Cambridge University Press.

Good, P. I. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. New York: Springer.

Härting, C., Markowitsch, H. J., Neufeld, H., Calabrese, P., Deisinger, K., & Kessler, J. (2000). *Die Wechsler-Memory-Scale-revised.* (German Adaptation). Bern: Huber.

Hildebrandt, A., Sommer, W., Herzmann, G. & Wilhelm, O. (submitted). Age differences in face cognition: Structural invariance and performance decline. *Psychology and Aging*.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Kline, R. B. (2005). *Principles and Practice of Structural Equation Modeling*. New York: The Guilford Press.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the LMS method. *Psychometrika, 65,* 457-474.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence, 14,* 389-433.

Labouvie, E., & Ruetsch, C. (1995a). Testing for equivalence of measurement scales: Simple structure and metric invariance reconsidered. *Multivariate Behavioral Research, 30,* 63-76.

Labouvie, E., & Ruetsch, C. (1995b). Wholes or parts? *Multivariate Behavioral Research, 30,* 121-123.

Lindenberger, U., & Baltes, P. B. (1997). Intellectual functioning in old and very old age: Cross-sectional results from the Berlin Aging Study. *Psychology and Aging, 12,* 410–432.

Little, T. D., Card, N. A., Slegers, D. W., & Ledford, E. C. (2007). Representing contextual effects in multiple-group MACS models. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 121–147). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Little, T. D., Slegers, D. W., & Card, N. A. (2007). A non-arbitrary method of identifying and scaling latent variables in SEM and MACS models. *Structural Equation Modeling, 13,* 59-72.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7,* 19-40.

McDonald, R. P. (1995). Testing for equivalence of measurement scales: A comment. *Multivariate Behavioral Research, 30,* 87-88.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543.

Meredith, W. (1995). Two wrongs may not make a right. *Multivariate Behavioral Research, 30,* 89-94.

Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide*. Fifth edition. Los Angeles, CA: Muthén and Muthén.

Nesselroade, J. R. (1995a). "…and expectation fainted, longing for what it had not." Comments on Labouvie and Ruetsch's "Testing for equivalence…". *Multivariate Behavioral Research, 30,* 95-99.

Nesselroade, J. R. (1995b). Further commentary on Labouvie & Ruetsch's "Testing for equivalence of measurement scales…". *Multivariate Behavioral Research, 30,* 119-120.

Oberauer, K., Schulze, R., Wilhelm, O., & Süß, H.-M. (2005). Working memory and intelligence – Their correlation and their relation: Comment on Ackerman, Beier, and Boyle (2005). *Psychological Bulletin, 131,* 61-65.

Oberauer, K., Süß, H-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity - facets of a cognitive ability construct. *Personality and Individual Differences, 29,* 1017-1045.

Preacher, K., J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods, 10,* 178-192.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), *Proceedings of the Forth Berkeley Symposium on Mathematical Statistics and Probability* (pp. 321-333). Berkeley, CA: University of California Press.

Raven, J. C., Court, J. H., & Raven, J. (1979). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. London: H. K. Lewis & Co.

Reinert, G. (1970). Comparative factor analytic studies of intelligence throughout the human life-span. In L. Goulet & P. Baltes (Eds.), *Life-span developmental psychology: Research and theory.* New York: Academic Press.

Schaie, K. W., Maitland, S. B., Willis, S. L., & Intrieri, R. C. (1998). Longitudinal invariance of adult psychometric ability factor structures across 7 years. *Psychology and Aging, 13,* 8–20.

Schaie, K. W., Willis, S. L., Jay, G., & Chipuer, H. (1989). Structural invariance of cognitive abilities across the adult life span: A cross-sectional study. *Developmental Psychology, 25,* 652–662.

Schmiedek, F., Hildebrandt, A., Lövden, M., Wilhelm, O., & Lindenberger, U. (2009). Complex span versus updating tasks of working memory: The gap is not that deep. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*, 1089-1096.

Schmiedek, F. (2005). Item response theory and the measurement of cognitive processes. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 393-408). Thousand Oaks, CA: Sage.

Schramm, U., Berger, G., Müller, R., Kratzsch, T., Peters, J., & Fröhlich, L. (2002). Psychometric properties of Clock Drawing Test and MMSE or Short Performance Test (SKT) in dementia screening in a memory clinic population. *International Journal of Geriatic Psychiatry, 17*, 254-260.

Schulze, R. (2005). Modeling structures of intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 241-263). Thousand Oaks, CA: Sage.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General, 125*, 4-27.

Small, B. J., Viitanen, M., & Bäckman, L. (1997). Mini-Mental State Examination item scores as predictors of Alzheimer's disease: Incidence data from the Kungsholmen project, Stockholm. *Journals of Gerontology, 52,* M299–M304.

Spearman, C. (1927). *The abilities of man.* London: Macmillan.

Stevens, S. S. (1951). Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York: Wiley.

Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurement.* New York: Teachers College, Columbia University.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology, 15,* 433-451.

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of mind.* Chicago: The University of Chicago Press.

Tucker-Drob, E. M., & Salthouse, T. A. (2008). Adult age trends in the relations among cognitive abilities. *Psychology and Aging, 23,* 453–460.

Tucker-Drob, E. M. (2009). Differentiation of cognitive abilities across the life span. *Developmental Psychology, 45,* 1097-1118.

Wu, H., & Zhang, J.-T. (2006). *Nonparametric regression methods for longitudinal data analysis.* Hoboken, NY: Wiley-Interscience.