Edinburgh Research Explorer

WAIS-III measurement invariance

WAIS-III measurement invariance: data from Estonian standardization

Kätlin Anni *

Institute of Psychology, University of Tartu, Estonia

Meelis Käärik

Institute of Mathematics and Statistics, University of Tartu, Estonia

René Mõttus

Department of Psychology, University of Edinburgh, UK

Institute of Psychology, University of Tartu, Estonia

* Corresponding author

Näituse 2, 50409, Tartu, Tartumaa, Estonia

Phone: +372 53 634 987

E-mail: katlinanni@gmail.com

Abstract

Objective: A prerequisite of any psychological instrument used to compare individuals from different groups is measurement invariance (MI). It indicates that the test measures the same psychological constructs regardless of the particular grouping variable of the test-taker. Our purpose was to evaluate the MI across sex, age groups and educational levels in the recently adapted Estonian version of the Wechsler Adult Intelligence Scale – Third Edition (WAIS-III). Method: We analysed the Estonian standardization sample of WAIS-III (N = 770) with confirmatory factor analysis (CFA) to establish the best baseline factor model for further analysis. Multi-group confirmatory factor analysis (MG-CFA) was applied to evaluate MI of the test and, granted this, mean differences across sex, age groups and educational levels. Results: CFA supported the four-factor model. The test demonstrated partial MI across sexes; latent mean comparisons showed that men had a significantly higher mean score on the Perceptual Organization factor. Partial MI also held across age groups and, as expected, older groups had significantly lower means than younger age groups. The analyses across the educational levels failed to prove the MI as the metric invariance was not tenable. Discussion: The results of this study provide evidence that the structural model underlying the Estonian adaption of WAIS-III is partially invariant across sex and age groups, hence the test functions same manner across these groups. Estonian WAIS-III was not invariant across the educational levels, which may indicate that the measure has a different structure or meaning to different educational groups.

Keywords: WAIS, measurement invariance, multi-group confirmatory factor analysis, sex differences, education, intelligence, neurocognitive test

WAIS-III measurement invariance: data from Estonian standardization

Psychological tests are used to measure a wide range of psychological variables for scientific purposes, but also to make practical decisions about individuals (Gregory, 2014). It is thus crucial that differences in test scores between people or groups are attributable to differences in the (underlying) properties that the particular test is developed to measure rather than something else (Borsboom, 2006). One of the fields that often rely on comparing test scores across groups is neuropsychology.

Measurement invariance

A prerequisite of any psychological instrument used to compare individuals from different groups (e.g., gender, age, educational level, health conditions) is measurement invariance (MI), which suggests that the test measures the same psychological constructs regardless of the particular grouping variable of the test-taker (Wicherts, 2016). MI is essential to establish not only for testing mean differences across groups, but also for comparing relations of the constructs with other variables across the groups (Putnick & Bornstein, 2016). It is only after establishing the MI that the interpretations of group comparisons are meaningful. Hence, the MI is among the central testing concepts in psychological, clinical and developmental sciences and an obligatory feature of any psychological measure (Putnick & Bornstein, 2016).

One of the most widely used methods to test for MI is multi-group confirmatory factor analysis (MG-CFA) (Milfont & Fischer, 2010; Putnick & Bornstein, 2016). In the confirmatory factor analysis (CFA) framework, observed indicators (e.g., items or subtests) which have been selected to measure an underlying construct are set to load on a latent factor that represent this ostensible construct. In a group comparison context, each indicator must relate to that latent variable in the same way across all the groups (Meredith, 1993). The

guidelines (Milfont & Fischer, 2010; Putnick & Bornstein, 2016; van de Schoot, Lugtig, & Hox, 2012) describe four steps for the assessment of MI with MG-CFA, which are based on Jöreskog's theoretical strategy (1971, 1993). The first step is evaluating the configural invariance, which involves testing whether the constructs have the same patterns of factor loadings across groups; essentially, if the same factors emerge in all groups. If configural invariance is tenable, the next step is to evaluate the metric (also known as weak) invariance, which means the equivalence of the items' loadings on the factors so that each item contributes to its latent variable to a similar degree across groups.

If the metric invariance is supported, the third step is to evaluate the scalar (strong) invariance, which tests the equivalence of the items' intercepts. This form of invariance ensures that the latent variable differences across groups are reflected in all indicators, proportionally to their factor loadings. If scalar MI is not met, then the observed group differences in scale scores are to some extent driven by the individual indicators of the latent trait rather than their shared variance (i.e., variance ostensibly due to the latent trait). This form of invariance is most commonly violated (e.g., Mõttus et al., 2015) and is also known as differential item functioning (Osterlind & Everson, 2009). The final step is to evaluate the residual (strict) invariance, which tests the equivalence of the residuals of the metric and the scalar invariant items. This form of invariance ensures that the latent variables are measured with the same degree of internal consistency (sometimes taken for reliability) across the groups.

Between every step, the differences between more restricted models (e.g., with loadings constrained equal across groups) and less restricted models (e.g., with loadings freely estimated in both groups) are examined. If the fit of the more restricted but more parsimonious model is significantly worse than that of the less parsimonious model, the tenability of the particular invariance step is not supported. As previous research has shown

that full MI in all four steps is rarely supported in practice, Byrne, Shavelson & Muthen

(1989) introduced the partial measurement concept. This means that some violations of

invariance are accepted by releasing the across-groups-equality constraints on one or more

loadings or intercepts, or both. It is suggested that more than half of the items of the

instrument should have invariant parameters across groups for meaningful comparisons to be

possible (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000).

Wechsler Intelligence Scales

The Wechsler Intelligence Scales, adapted to many countries, are among the most

widely used intelligence tests in scientific research as well as in clinical practice. The

Wechsler Adult Intelligence Scales (WAIS), Wechsler Intelligence Scales for Children

(WISC), Wechsler Preschool and Primary Scale of Intelligence (WPPSI) and Wechsler

Abbreviated Scale of Intelligence (WASI) have gone through several updates to incorporate

theoretical advances in intelligence conceptualizations as well as advances in psychometrics,

neuropsychology and cognitive neuroscience (Coalson, Raiford, Saklofske, & Weiss, 2010).

The WAIS and WPPSI are in their fourth edition and WISC in its fifth edition, whereas WASI

has had two editions. Essentially, these tests are the golden standard of intelligence

measurement. In the current study, we focused on the Wechsler Adult Intelligence Scale –

Third Edition (WAIS-III; Wechsler, 1997), the only Wechsler scale adapted to Estonia yet.

The WAIS-III had many updates compared to its preceding editions. One of the most

important updates was the addition of new subtests, which provide reliable scores for four

cognitive domains – Verbal Comprehension, Perceptual Organization, Working Memory and

Processing Speed – in addition to the general IQ scores. The intended four-factor structure

was indeed confirmed by CFA for the original US version of WAIS-III (Psychological

Corporation, 2002; Tulsky & Price, 2003), although studies have also discussed the merits of

two- and three-factor models (Kaufman, Lichtenberger, & McLean, 2001; Ward, Ryan, &

Axelrod, 2000) or the hierarchical models based on the Cattell–Horn–Carroll (CHC) theory (Golay & Lecerf, 2010). Several studies have demonstrated the robustness of the four-factor structure, including a re-analysis of original US data (Deary 2001) as well as the Canadian adaptation (Bowden, Lange, Weiss, & Saklofske, 2008) and adaptations into other languages (Egeland et al., 2009; Grégoire, 2004). Although a more recent edition of WAIS has been published (WAIS-IV; Wechsler, 2008), it has maintained the similar structure of four domains; this allows researchers and practitioners alike to compare the results from two WAIS editions.

It has become a common practice to thoroughly evaluate the psychometric properties such as the reliability and validity of updated or adapted tests, so as to make sure that the interpretation of results remains valid. Wicherts (2016) also highlights the need to test the MI of the scales across commonly assessed groups, especially when adapting tests or collecting appropriate normative data. However, MI tests based on CFA often fail in commonly used neurocognitive batteries (Wicherts, 2016). For example, some studies have failed to show MI in the scales across ethnics groups, while others have only been able to demonstrate partial MI (Dolan, Roorda & Wicherts, 2004; Wicherts & Dolan, 2010). Also, measurements across sex, age groups or normal vs clinical samples have been non-invariant or only partially invariant (Chen & Zhu, 2012; Dolan et al., 2006; Niileksela, Reynolds, & Kaufman, 2013). Such inconclusive results highlight the need for further invariance studies, because they show that the scores of these tests may often be differentially affected by some population groupings and not only by the latent cognitive abilities that the tests have been developed to measure.

The Wechsler scales are regularly used to compare gender or educational groups for research purposes, healthy controls and patients in clinical psychology, neuropsychology, rehabilitation services or forensic contexts, and to compare client groups in educational and counselling services. These are just some of the many applications of the scales. Therefore,

establishing the MI of adapted norms across these groups is particularly important (Chen, Zhang, Raiford, Zhu & Weiss, 2015; Millsap & Kwok, 2004; Wicherts, 2016). And yet, Wicherts (2016) claimed that the importance of MI may be under-appreciated and it should be routine procedure for assessing the adequacy of norms in neurocognitive measures.

In response to this, our aim was to evaluate the MI across sex, age and educational levels in the recently adapted Estonian version of Wechsler Adult Intelligence Scale – Third Edition (WAIS-III; Wechsler, in press). Several MI studies have focused on MI across sex - perhaps the most common population grouping - and have found WAIS-III to be partially invariant, because results did not support the full metric (Dolan et al., 2006) or scalar invariance (van der Sluis et al., 2006). This means that there were sex differences on subtests that could not be explained by sex differences on the relevant domain scores of WAIS-III - the ostensible latent traits underlying cognitive performance in specific subtests. For example, Arithmetic and Information showed larger sex differences in favour of males, controlling for the factor loadings of these tests on their domains. It means that the differences between males and females were larger than would be expected on the basis of any potential sex differences on the corresponding latent traits. On the other hand, Chen and colleagues (2015) reported invariance across genders with WISC-IV. It could indicate that sex differences of the measured constructs are most pronounced in young adulthood (Lynn & Irwing, 2004), but it also confirms the need to study invariance with different test versions and contexts.

MI across age is also critical for attempts to establish how age is associated with cognitive ability and its changes. Testing MI can assure that the underlying structure of the specific test is stable across a range of ages. Although the Wechsler tests have not originally been developed according to a specific underlying theory, it is somewhat surprising that the replicated four-factor structure may be invariant even across a wide age range (Bowden, Weiss, Holdnack & Loyd, 2006; Taub, McGrew & Witta, 2004).

The level of education, another major dimension of population stratification, is also an important variable for the MI assessment. A strong association between intelligence test performance and educational levels is established in many studies (Strenze, 2007). However, if the cognitive measures are not invariant with respect to educational attainment, then the subtests might show bias for specific groups; in other words, the differences may be driven by some subtests rather than a general intelligence *per se*. There have been contradictory results regarding the MI across educational level. Tommasi et al. (2015) studied the Italian standardization sample of WAIS-R and concluded that the MI was tenable, whereas Abad, Sorrel, Román, and Colom (2015) found partial metric MI for the Spanish WAIS-IV: three subtests (Matrix Reasoning, Coding and Letter-Number-Sequencing) showed lower loadings as the educational level increased, whereas scalar invariance and strict invariance were supported. These results indicate that the comparability of test results between educational groups cannot be taken for granted – more research is needed.

The current study begins with identifying the factor structure of the Estonian adaptation of WAIS-III to specify the basic model for MI. As no Estonian WAIS-III factor analysis had been published earlier, we could, for the first time, compare the factor structure of an Estonian WAIS to that of the test's original version and to those of other adaptations such as French (Grégoire, 2004), Norwegian (Egeland, Bosnes, & Johansen, 2009) and Canadian (Saklofske, Hildebrand, & Gorsuch, 2000), among others. Our aim was to evaluate the factor models that have been tested most frequently in previous research. Among these, the intended four-factor model has been particularly well established for WAIS-III (Psychological Corporation, 2002; Tulsky & Price, 2003), hence, our hypothesis was that a similar four-factor solution will be confirmed in the Estonian data. Subsequently, MI across gender, age groups and educational levels was tested. Where the MI was supported to a sufficient degree across these groups, it enabled us to compare their mean scores.

Method

Sample

We analysed the Estonian standardization sample of WAIS-III, which was stratified by age, gender and educational level following the same exclusion criteria used by the original WAIS-III standardization (Psychological Corporation, 2002). The final sample consisted of 770 subjects (341 males, 429 females) and its composition was adjusted to the theoretical percentages based on the Estonian census data (the final composition plan was renewed during data collection in 2014). All participants gave written informed consent to take part in the study and did not receive any compensation. Ethical approval was granted by the Ethic Review Committee on Human Research, University of Tartu, Estonia.

The mean age of males was 41.04 years (SD = 19.83) and the mean age of females was 48.42 years (SD = 22.39); this difference was medium in size and statistically significant, $t(768)$ = -4.78, p < .001, Cohen's d = .349. Among male participants, 31.7% had basic education, 54.5% had secondary or vocational education and 13.8% had higher education, whereas among females 25.9% had basic education, 49.7% had secondary or vocational education and 24.5% had higher education. The proportions of educational levels across gender groups were significantly different, $\chi^2(2, N = 769)$ = 14.12, p < .001. The effect size (Cramer's V = .095) for this analysis can be considered small to medium (Cohen, 1988).

For the purpose of sufficiently sized age groups for MI analyses, we divided the sample to three groups: 16-29 years (N = 242), 30-54 years (N = 252) and 55-89 years (N = 276). Educational levels of the youngest age group (16-29 years) were the following: 52.5% had basic education, 36.4% secondary or vocational education and 11.2% higher education. The proportions for educational levels of age group 30-54 years were the following: 15.9% had basic education, 60.7% had secondary or vocational education and 23.4% had higher education. The educational levels of age groups 55-89 years were: 18.8% basic education,

57.2% secondary or vocational education and 23.4% higher education. The proportions of educational levels across age groups were significantly different, χ^2 (4, N = 769) = 101.87, p < .001. The effect size (Cramer's V = .182) for this analysis can be considered medium to large (Cohen, 1988).

For the MI analysis across educational levels, we limited the sample by age, because we assumed that the younger participants might be still in process of attaining education and the oldest participants' opportunities for education may have been somewhat restricted. Indeed, an analysis of variance (ANOVA) revealed a significant effect of the age group on educational attainment, F(10, 759) = 23.9, p < .001 with full sample. Post hoc analyses showed significant mean educational level differences for age groups 16-17 and 18-19 *(p* < .001). In the age range of 20-89, the effect of age was not significant according to the ANOVA, F(2, 660) = 1.083, p = .339 (effect size partial η^2 = .003). The final sample for analyses across educational levels was 663 participants in the age range of 20-89, which can be considered more homogeneous on mean educational level. During the norming studies, we distinguished several educational levels, but for this study, we composed three larger educational groups – basic level (basic and primary school, up to 9 years of education; N = 122), secondary level (secondary and vocational education, 10-12 years of education; N = 389) and higher level (higher education, 13-20 years of education; N = 152). The mean ages of groups with basic, secondary and higher educational levels were 50.7 years, 48.7 years and 51.2 years, respectively.

The rest of the analyses (across genders and across age groups) were performed using the full sample.

Measures

The standardization of WAIS-III in Estonia was completed in 2019 (Wechsler, in press). The WAIS-III contains 14 subtests, which provide a Full Scale IQ, a Verbal IQ and a Performance IQ. It also provides four index score factors: Verbal Comprehension (Vocabulary, Similarities, Information, Comprehension subtests), Perceptual Organization (Picture Completion, Block Design, Matrix Reasoning), Working Memory (Arithmetic, Digit Span, Letter-Number Sequencing), and Processing Speed (Digit Symbol – Coding, Symbol Search). The normative data for subtests was developed using the inferential norming method (Zhu & Chen, 2011).

The Estonian adaptation of WAIS-III has mostly acceptable to excellent internal consistency statistics that is comparable with the original version (Psychological Corporation, 2002). The average reliability coefficients (Cronbach's alphas) across 11 age groups were .97 for Full Scale IQ, .96 for Verbal IQ and .92 for Performance IQ.

Statistical analyses

The analyses were based on raw subtest scores. The Object Assembly subtest was not included, because it is an optional subtest often left unanalyzed in previous studies (e.g. Bowden et al., 2006; Egeland et al., 2009; Grégoire, 2004; Tulsky & Price, 2003), whereas one of our aims was to compare our results with previous research.

Confirmatory factor analyses to select the best baseline model. We applied the confirmatory factor analysis to the data to identify the factorial structure of the Estonian WAIS-III. We tested nine models that have been studied in prior research with the original scale (Psychological Corporation, 2002) and previous adaptations (Egeland et al., 2009; Grégoire, 2004; Tulsky & Price, 2003). As newer editions of Wechsler's scales have been

also analysed according to the CHC framework (McGrew, 2009), we tested models based on that as well. We compared the goodness-of-fit statistics for the following models:

1. Model 1: A one-factor model that includes one general *g*-factor underlying all of the 13 subtests.

2. Model 2: A two-factor model, which corresponds to the traditional organization of the Wechsler scales into a Verbal and Performance scale (Verbal IQ = seven verbal subtests, Performance IQ = six performance subtests).

3. Model 3: A four-factor model as suggested in the Technical Manual of the original WAIS-III version (Psychological Corporation, 2002). The factors are Verbal Comprehension (VC = Vocabulary, Similarities, Information, Comprehension), Perceptual Organization (PO = Picture Completion, Matrix Reasoning, Block Design, Picture Arrangement), Working Memory (WM = Arithmetic, Digit Span, Letter-Number Sequencing) and Processing Speed (PS = Coding, Symbol Search).

4. Model 3a: A four-factor model where Arithmetic loads on the VC factor instead of the WM factor, as proposed by Egeland et al. (2009); the factors were allowed to correlate.

5. Model 3b: A four-factor model where Arithmetic is allowed to load on both the VC and the WM factors, as suggested by Egeland et al. (2009) and Tulsky and Price (2003); the factors were allowed to correlate.

6. Model 4: A hierarchical model with four first-order factors (same as in Model 3a) but with a second-order general factor.

7. Model 4a: A hierarchical model with four first-order and a second-order general factor, but with Arithmetic allowed to load on both the VC and the WM factors (similarly as in Model 3b).

8. Model 5: A model based on the CHC framework with five first-order factors and a second-order general factor g. The factors are crystallized intelligence factor (Gc = Vocabulary, Similarities, Information, Comprehension), visual processing factor (Gv = Picture Completion, Block Design, Picture Arrangement), fluid reasoning factor (Gf = Matrix Reasoning, Arithmetic), short-term memory factor (Gsm = Digit Span and Letter-Number Sequencing) and processing speed factor (Gs = Coding, Symbol Search).

9. Model 5a: A model based on the CHC framework with five first-order factors and a second-order general factor g as Model 5, but the Arithmetic is allowed to load on both the Gf and the Gsm factors.

Following the goodness-of-fit indices were considered when evaluating the fit of the factor model: the chi-square (χ^2), the comparative fit index (CFI), the Tucker−Lewis fit index (TLI), the root mean square error of approximation (RMSEA) and the 90% confidence interval for RMSEA. A good model should have CFI ≥ .95, TLI ≥ .95, RMSEA ≤ .06, an inferior limit of the 90% RMSEA confidence interval ≤ .08, and an acceptable model should have CFI and TLI ≥ .90 and RMSEA ≤ .08 (Browne & Cudeck, 1993; Hu & Bentler, 1999). The models were identified by fixing the variance of latent variables at unity.

Invariance analyses. MG-CFA was applied to test for MI based on a set of nested models (Milfont & Fischer, 2010; Putnick & Bornstein, 2016; van de Schoot, Lugtig, & Hox, 2012):

1. The baseline configural invariance model, with loadings and intercepts free to vary across specific grouping variables, but the same factorial pattern was specified for each group; means were constrained equal.

2. The metric invariance model, with loading constrained to be equal across specific grouping variables; means were constrained equal.

3. The scalar invariance model, with factor loadings and intercepts constrained to be equal

across grouping variables.

4. The strict invariance model, with factor loadings, intercepts and residual variances

constrained to be equal across grouping variables.

The difference between CFIs (ΔCFI) of invariance models was estimated for testing

the MI. Cheung and Rensvold (2002) propose that the ΔCFI is one of the best indices to test

MI, because it is unaffected by sample size and model complexity, unlike the chi-square

difference ($\Delta\chi^2$) test. A value of ΔCFI (more constrained model minus less constrained

model) smaller than or equal to -.002 indicates that the null hypothesis of invariance should

not be rejected (Meade, Johnson, & Braddy, 2008).

All the statistical analyses were conducted using the R Statistical software (R

Development Core Team, 2018); for confirmatory factor analysis the *lavaan* package

(Rosseel, 2012) was used.

Results

Confirmatory factor analysis of Estonian WAIS-III

We tested Models 1 to 5a as single group models to choose the most appropriate

baseline model for further MI analysis.

A correlated four-factor model (Model 3) provided the best overall fit to the data,

when we compared it with one- or two-factor models (Table 2). However, as the RMSEA was

.070, we modified the model further. Modification indices indicated that the model would

improve if the Arithmetic subtest would load on the VC factor instead of the WM factor

(Model 3a), but the fit indices did not improve. Again, based on the modification indices and

previous research, we allowed the Arithmetic subtest to load both on the VC and the WM factors. The fit indices of this model (Model 3b) showed improvement (Table 1).

Next, we also tested hierarchical models with four first-order factors (VC, PO, WM, PS) and one second-order factor (*g*); these are Models 4 and 4a. Fit indices of the hierarchical models showed no improvement compared to the first-order models (Models 3 and 3b); again allowing the Arithmetic subtest to load both the VC and the WM factors resulted in a better fit.

Comparing the first-order Model 3b to the hierarchical Model 4a, the fit indices were better for the first-order model. As it is also a longstanding tradition to favour simpler models over a more complex model (Bollen & Long, 1993), we decided to use the Model 3b in further MI analyses. Path diagram with the standardized factor loadings and covariances between factors are shown in Supplemental material.

MI across sex

Table 2 shows the means and standard deviations of the subtest scores and composite scores for males and females. Effect sizes (Cohen's *d*) are calculated as the differences between the means for males and females divided by their pooled standard deviation. According to Cohen's (1988) recommendations, effect sizes .20 can be interpreted as small, .50 as medium and .80 as large.

Fit indices for MI testing between males and females are shown in Table 3. The configural invariance was satisfied, with CFI (> .95) and RMSEA (< .06) indicating a reasonable fit. The metric invariance was also tenable as the equality of subtest loadings did not result in a significant degradation of model fit (ΔCFI < .002). However, constraining intercepts equal (for scalar invariance) did yield a significant degradation of model fit (ΔCFI = .014). We examined the modification indices of the model and sequentially released

intercept constraints according to the suggestions from Yoon & Kim (2014). We retested the model until the model degradation criterion was achieved (Table 3). Sequentially releasing equality of intercepts for Information (Model 3a), Arithmetic (Model 3b) and Coding subtests (Model 3c) resulted in an acceptable difference between Model 3c and Model 2 (ΔCFI $\leq$ .002). We discontinued the analysis and did not test strict variance by constraining residual variances because only partial scalar invariance was tenable.

As most of the intercepts in factors remained constrained, we tested the sex differences regarding the four latent factors VC, WM, PS and PO. Table 4 shows the differences of latent factors in a partial MI model, where the Information, Arithmetic and Coding subtests' intercepts were released. Factor means were fixed to zero in females, whereas means of the males were estimated as a deviation of the mean of the females (the unit of variance was standard deviation, so the effects are in Cohen's d metric). Males outperformed females in the PO factor ($d = .369$), whereas there were no significant sex differences in the other factors.

MI across age groups

Descriptive data regarding age groups are shown in Table 5. Fit indices for MI analyses across age groups are shown in Table 6. The configural invariance was met, with CFI ($> .95$) and RMSEA ($< .06$) indicating a good fit. Metric invariance was not tenable as imposing the equality on subtest loadings across groups resulted in a significant degradation of model fit (ΔCFI $= .006$). We examined the modification indices and sequentially released the constraints from Block Design subtest loading to PO factor and Martrix Reasoning loading to PO factor (Table 6). Releasing constraints from both of the loadings lowered the model degradation to our criterion of the MI (ΔCFI $\leq .002$) and partial metric invariance was thus confirmed. Scalar invariance was tested by constraining the item intercepts to be equivalent across groups for metric invariant items (the loadings of Block Design and Matrix Reasoning subtests allowed to vary). Fit indices showed that scalar invariance was not

tenable, as the degradation of model fit was significant (ΔCFI = .013). We investigated the source of the noninvariance by sequentially releasing item intercept constraints and we retested the model until a partially invariant model was confirmed. Intercepts of Picture Arrangement, Arithmetic, Vocabulary and Information subtests needed to be released to achieve an acceptable difference between Model 2b and Model 3d (ΔCFI < .002). We discontinued the invariance testing because full strict measurement invariance was unachievable.

We decided to compare the differences of latent factors (VC, WM, PS and PO) because at least half of the items in every factor were constrained in the invariance analyses. Factor means were fixed to zero in the youngest age group, whereas the factor means of the two older age groups were estimated as a deviation of the youngest one. Hence, the negative means can be interpreted as downward age trends and the positive means as upward age trends (again, in Cohen's *d* metric). As expected, the older age groups had mostly lower means. The age group of 30-54 years had a slightly higher latent mean in the VC compared to the youngest age group, but the difference was not statistically significant.

MI across educational levels

Descriptive data are shown in Table 8 and fit indices for MI analyses are shown in Table 9. The configural invariance was confirmed with a reasonable fit (CFI > .95; RMSEA < .06). The metric invariance was not tenable as the equality of subtest loadings resulted in a significant degradation of model fit (ΔCFI = .007). We investigated the source of noninvariance and an examination of the modification indices revealed that the model would improve if loading of Information to VC was released. This adjustment was not sufficient, but further examination of the modification indices showed that the largest remaining modification index was not statistically significant; hence, the model could not be improved.

This means that the partial measurement invariance was not tenable as well and we discontinued further invariance testing.

Discussion

In the present study, we analysed the MI of Estonian WAIS-III across sex, age and educational levels, using the multi-group confirmatory factor analysis.

First, we established the baseline factor structure of the Estonian WAIS-III. The results of the confirmatory factor analyses of the standardization sample supported the four-factor model, with the latent factors of Verbal Comprehension, Perceptual Organization, Working Memory and Processing Speed. These results replicate the solution found in the original version of WAIS-III (Wechsler, 1997; see also Saklofske et al., 2000) as well as subsequent standardizations (Egeland et al., 2009; García, Ruiz, & Abad, 2003; Grégoire, 2004). The fit of models with one or two factors as well as models based on CHC framework was inadequate. Documenting the factor structure in different adaptations and standardization samples is important because it allows to establish the universality of the underlying structure of the scale and thereby cognitive abilities more generally, beside the validity of the adaptation (especially in the case of WAIS-III, which is so commonly used in many adaptations).

Egeland et al. (2009) reported of a Norwegian sample that in the most parsimonious model of WAIS-III, the Arithmetic subtest did not load to one single factor, but had to be allowed to load on both WM and VC factor. The same also appeared in the current study, which confirms that the Arithmetic subtest is somewhat multifaceted. Changing the Arithmetic subtest loading from the initial WM factor to the VC factor somewhat increased the fit indices, but the best fit appeared when the subtest was allowed to load on both factors. A possible interpretation of this is that the Arithmetic subtest is composed of word problems

that require verbal comprehension to give the right answers (Arnau & Thompson, 2000), but solving these problems needs a broader working memory involvement as well (Tulsky & Price, 2003). Egeland et al. (2009) also found that education explained a larger part of the variance in the VC subtests and in the Arithmetic subtest, but less in the other WM subtests. Similar issues with the Arithmetic subtest have also been pointed out by other authors – for example, the factor analysis studies in France (Grégoire, 2004) and the re-analysis of the original scale (Arnau & Thompson, 2000; Tulsky & Price, 2003). Therefore, we used a four-factor solution with splitted loading on the VC and the WM in further MI analyses.

The main aim of this study was to evaluate the MI across different groups. The results show that the Estonian WAIS-III has a partial MI across sexes. The configural and metric invariances were satisfied, whereas the scalar invariance was tenable only after the intercepts of Information, Arithmetic and Coding were released. According to the descriptive data and the comparison of observed means, males were found to outperform females on two of the 14 subtests – Information and Arithmetic – and females outperformed males on two processing speed subtests: Coding and Symbol Search. The MI analysis showed the similar results that Information, Arithmetic and Coding subtests were biased, so we allowed its intercepts to vary freely across genders when comparing latent factor means. Males and females showed no mean differences of the factors VC, WM or PS. However, males had a significantly higher mean score of the PO factor. Even so, it is questionable if the latent means were in fact comparable because of the partial invariance. There are no universal recommendations for how the partially invariant models influence the accuracy of mean-level comparisons (Putnick & Bornstein, 2017). Steinmetz (2013) found that the effects of scalar noninvariance might be large. More research is definitely needed, as there are no clear solutions how to manage the partial noninvariance (Putnick & Bornstein, 2017), although in practice partial subtest intercept invariance is not uncommon (Immekus & Maller, 2010).

These results are in concordance with the previous studies. The analysis of sex differences on Dutch (van der Sluis et al., 2006) and Spain (Dolan et al., 2006) WAIS-III revealed a similar pattern of differences. In both studies, men had higher scores in the Information subtest. This finding is well documented with several previous studies with the Information subtest of the Wechsler scales or similar overall general knowledge tests (Lynn, Irwing, & Cammock, 2001; Lynn, Wilberg, & Margraf-Stiksrud, 2004). Recent meta-analysis (Tran, Hofer, & Voracek, 2014) also found some male advantage of general knowledge, but their analysis indicated that these sex differences could be explained by the differences in schooling and selection processes that were moderated by the parental education.

Similar to the current study, analysing latent factor means, van der Sluis et al. (2006) and Dolan et al. (2006) found no sex differences in the VC factor and that males outperformed females in the PO factor. The absence of sex differences in verbal ability have been found in earlier studies as well, which have not specifically looked for MI (see further the meta-analysis by Hyde and Linn, 1988). Comparable to our findings, females did not show any advantages over males in the PS in a Spanish study (Dolan et al., 2006), although Dutch females outperformed males in the PS factor (van der Sluis et al., 2006). Both studies also found that males outperformed females in the Working Memory factor, which was not the case with the present study. Gender differences are therefore possibly culture-specific.

Next, we analysed the MI across three age groups and concluded that partial MI is tenable. The configural invariance was satisfied. The metric and scalar invariances were not entirely tenable and the constraints of some subtests needed to be released to result in an acceptable model fit. We released loadings of the Block Design and Matrix Reasoning subtests to the PO factor as testing the metric invariance. We released the intercepts of the Picture Arrangement, Arithmetic, Vocabulary and Information subtests as testing the scalar invariance. None of the items had full noninvariance with both the loading and the intercept

being released. Again, it is questionable how releasing constraints would influence the mean difference analysis. However, as most of the items in factors were constrained we explored the differences between latent factors. The results were as expected with the lowest means in the older age groups, the largest discrepancies in the PS factor and the smallest differences in the VC factor.

The finding that the MI in most part held across the age groups is significant in many ways. It ensures that the measure is comparably usable both in the younger age groups as well as in the older age groups, which has a critical value for diagnostic or classification purposes. The MI also shows that the underlying constructs are stable across the age groups, which is an important property for both the psychological constructs themselves and their test (Bowden et al, 2006). In the case of many degenerative diseases there is a need to conduct repeated assessments, often over extended retest intervals (Horn & McArdle, 1992), so the MI is crucial to adequately interpret the changes across aging (Bowden et al., 2006). However, the MI across age groups is also relevant for the very concept of intelligence. For example, it has been argued that if intelligence factors such as g emerge developmentally as a consequence of mutually beneficial interactions among the specific skills (dynamic mutualism approach by van der Maas et al., 2006), their co-variances should *not* be structurally invariant. In response to this theory, Gignac (2014) tested the mutualism approach and g models with various Wechsler scales and his results did not support the mutualism model, because the g factor was present and constantly strong across the development.

We further established that the Estonian WAIS-III is not invariant and thereby likely to be biased across educational levels. Measurement noninvariance means that the construct has a different structure or meaning to different groups (Putnick & Bornstein, 2016). In turn, group mean differences cannot be interpreted in terms of the latent cognitive abilities (Wicherts, 2016). According to Wicherts (2016), the failure of MI with respect to some

subgroups in the standardization sample would raise a question whether it is appropriate to use overall norms. Wicherts (2016) proposes that a possible solution for the noninvariance of cognitive tests may be to develop subgroup norms or to revise the subtests (adaptations) to correct the bias.

To our knowledge, MI analyses of WAIS-III across educational levels have not been previously published, although some recent results of the WAIS-R and WAIS-IV invariance are available. Tommasi et al. (2015) found the MI across educational levels tenable with the WAIS-R, while Abad et al. (2015) recently studied the invariance across educational levels with the WAIS-IV sample from Spain. They concluded that the factor structure of the WAIS-IV was only partially invariant, as three subtests (Matrix Reasoning, Coding and Letter-Number Sequencing) showed lower loadings as the educational level increased. The differences between these previous studies and our study may stem from various causes. Firstly, different editions of the Wechsler Scales are similar, but not exactly the same, so the structure may depend on the changes made throughout subsequent versions. Secondly, the results may be influenced by the language/location, where the test was adapted and the sample was collected. Besides language, the differences may be in the composition of samples, divisions of the educational levels and differences based on the overall educational system. Therefore, it is crucial to study the relationships between the different models, theories of intelligence structure, educational systems and locations more widely to make further conclusions. Wicherts and Dolan (2010) have discussed additional reasons for intercept differences in the intelligence test CFA models, for example test-taking strategies, familiarity with testing in general and tests in particular or abilities that are tapped by certain subtest and that are distinct from the targeted latent ability.

Some limitations of our study deserve attention. A larger sample would add power to the analyses. The sizes of the groups divided by educational level were somewhat uneven, for

example the sample with basic education had 121 participants, while the group with

secondary education had 389 participants. The sample composition was based on the Estonian

population and we controlled that the different age groups did not differ significantly by

educational level, which may provide a partial solution to this problem. Second, as our overall

sample was already small, we differentiated the education only by three levels, which allowed

the groups to be sufficient in size. Another division of more specific educational paths may

have given different results, although a more complex study design with a larger sample size

and equal groups is needed to investigate these issues further. In addition, significant

differences in the demographic characteristics between groups may have influenced the

results. We found small to medium effect sizes for the analyses of age and education

differences between genders. The effect size was medium to large when comparing education

differences across age groups.

Future studies would benefit from the MI analyses with the clinical samples as well,

especially if the MI with standardization sample is tenable and proves the validity of the

measure. As the neurocognitive measures are often used with clinical populations, it is crucial

to make sure that the factor structure proposed with a normative sample will be supported in

various other diagnoses. There is some evidence that the MI for neurocognitive tests may not

hold, when patients are compared to healthy controls (Haring, Mõttus, Koch, Trei & Maran,

2014). In addition, evaluating the MI between ethnic groups could also reduce the possibility

of bias in mental testing. For example, it has been claimed that mean differences between

racial or ethnic subgroups result from problems in the construction, design or interpretation of

tests, not from real group differences in the ability (Brown, Reynolds, & Whitaker, 1999).

In conclusion, the results of this study provide evidence that the structural model

underlying the Estonian adaptation of WAIS-III is partially invariant across sex and age

groups but not invariant across educational levels. Our study also presents the additional

information on the sex differences of cognitive ability in Estonia. As Wicherts (2016) pointed out, assessment of the MI provides a way to empirically test whether tests of the cognitive ability measures function in the same manner across the different groups. We can conclude that the results of the current study provide some evidence of the appropriateness of the Estonian WAIS-III normative data, but the reasons of noninvariace across the educational levels needs to be studied further as it was not in the scope of this study.

Disclosure statement

The authors declare no conflicts of interest.

Funding

References

Abad, F. J., Sorrel, M. A., Román, F. J., & Colom, R. (2015). Relationships between WAIS-IV factor index scores and educational level: A bifactor model approach. *Psychological Assessment, 28*(8), 987-1000. doi:10.1037/pas0000228

Arnau, R. C., & Thompson, B. (2000). Second order confirmatory factor analysis of the WAIS-III. *Assessment, 7*(3), 237-246. doi: 10.1177/107319110000700304

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Scott (Eds.), *Testing Structural Equation Models* (pp. 1-9). Sage: Newbury Park, CA.

Borsboom, D. (2006). When does measurement invariance matter? *Medical Care, 44*(11), 176-181. doi:10.1097/01.mlr.0000245143.08679.cc

Bowden, S. C., Lange, R. T., Weiss, L. G., & Saklofske, D. H. (2008). Invariance of the measurement model underlying the Wechsler Adult Intelligence Scale–III in the United States and Canada. *Educational and Psychological Measurement, 68*(6), 1024-1040. doi: 10.1177/0013164408318769

Bowden, S. C., Weiss, L. G., Holdnack, J. A., & Lloyd, D. (2006). Age-related invariance of abilities measured with the Wechsler Adult Intelligence Scale-III. *Psychological Assessment, 18*(3), 334-339. doi:10.1037/1040-3590.18.3.334

Brown, R. T., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental testing since bias in mental testing. *School Psychology Quarterly, 24*(3), 208-238. doi: 10.1037/h0089007

Browne, M. W. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

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.

    Psychological Bulletin, 105, 456-466.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance.

    Structural Equation Modeling: A Multidisciplinary Journal, 14(3), 464-504. doi:

    10.1080/10705510701301834

Chen, H., Zhang, O., Raiford, S. E., Zhu, J., & Weiss, L. G. (2015). Factor invariance

    between genders on the Wechsler Intelligence Scale for Children-Fifth edition.

    Personality and Individual Differences, 86, 1-5. doi:10.1016/j.paid.2015.05.020

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing

    measurement invariance. *Structural Equation Modeling, 9*(2), 233-255. doi:

    10.1207/S15328007SEM0902_5

Coalson, D. L., Raiford, S. E., Saklofske, D. H., & Weiss, L. G. (2010). WAIS-IV: Advances

    in the assessment of intelligence. In L. G. Weiss, D. H. Saklofske, D. L. Coalson, &

    S. E. Raiford (Eds.), *WAIS-IV clinical use and interpretation* (pp. 3-23). San Diego,

    CA: Academic Press. doi:10.1016/C2009-0-01910-2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.).* Hillsdale,

    NJ: Lawrence Erlbaum Associates.

Dolan, C. V., Colom, R., Abad, R. J., Wicherts, J. M., Hessen, D. J., & van de Sluis, S.

    (2006). Multi-group covariance and mean structure modeling of the relationship

    between the WAIS-III common factors and sex and educational attainment in Spain.

    Intelligence, 34, 193-210. doi:10.1016/j.intell.2005.09.003

Egeland, J., Bosnes, O., & Johansen, H. (2009). Factor structure of the Norwegian version of the WAIS-III in a clinical sample: the arithmetic problem. *Assessment, 16*(3), 292-300. doi:10.1177/1073191108324464

García, L. F., Ruiz, M. A., & Abad, F. J. (2003). Factor structure of the Spanish WAIS-III. *Psicothema, 15*, 155-160.

Gignac, G. E. (2014). Dynamic mutualism versus g factor theory: An empirical test. *Intelligence, 42*, 89-97. doi:10.1016/j.intell.2013.11.004

Golay, P., & Lecerf, T. (2011). Orthogonal higher order structure and confirmatory factor analysis of the French Wechsler Adult Intelligence Scale (WAIS–III). *Psychological Assessment, 23*(1), 143-152. doi: 10.1037/a0021230

Grégoire, J. (2004). Factor structure of the French version of the Wechsler Adult Intelligence Scale-III. *Educational and Psychological Measurement, 64*(3), 463-474. doi:10.1177/0013164403258452

Gregory, R. J. (2014). *Psychological testing: History, principles and applications, 7th edition*. Boston, MA: Pearson.

Haring, L., Mõttus, R., Koch, K., Trei, M., & Maron, E. (2015). Factorial validity, measurement equivalence and cognitive performance of the Cambridge Neuropsychological Test Automated Battery (CANTAB) between patients with first-episode psychosis and healthy volunteers. *Psychological Medicine, 45*, 1919-1929. doi:10.1017/S0033291714003018

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3-4), 117-144. doi: 10.1080/03610739208253916

Hu, L., & Bentler, P. M. (2009). Cutoff criteria for fit indexes in covariance structure

analysis: Conventional criteria versus new alternatives. *Structural Equation

Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi:10.1080/10705519909540118

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis.

Psychological Bulletin, 104(1), 53-69.

Immekus, J. C., & Maller, S. J. (2010). Factor structure invariance of the Kaufman

Adolescent and Adult Intelligence Test across male and female samples. *Educational

and Psychological Measurement, 70*(1), 91-104. doi:10.1177/0013164409344491

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika

36*, 409-426.

Jöreskog, K.G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long

(Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage

Publications.

Kaufman, A. S., Lichtenberger, E. O., & McLean, J. E. (2001). Two- and three-factor

solutions of the WAIS-III. *Assessment, 8*(3), 267-280. doi:

10.1177/107319110100800303

Lynn, R., & Irwing, P. (2004). Sex differences on the progressive matrices: A meta-analysis.

Intelligence, 32, 481-498. doi:10.1016/j.intell.2004.06.008

Lynn, R., Irwing, P., & Cammock, T. (2002). Sex differences in general knowledge.

Intelligence, 30, 27-39. doi:10.1016/S0160-2896(01)00064-2

Lynn, R., Wilberg, S., & Margraf-Stiksrud, J. (2004). Sex differences in general knowledge in

German high school students. *Personality and Individual Differences, 37*, 1643-

1650. doi:10.1016/j.paid.2004.02.018

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the

    shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1-10.

    doi:10.1016/j.intell.2008.08.004

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative

    fit indices in tests of measurement invariance. *The Journal of Applied Psychology,*

    93(3), 568-92. doi:10.1037/0021-9010.93.3.568

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance.

    Psychometrika, 58, 525-543. doi:10.1007/BF02294825

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups:

    Applications in cross-cultural research. *International Journal of Psychological*

    Research, 3(1), 111-121. doi: 10.21500/20112084.857

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on

    selection in two populations. *Psychological Methods, 9*, 93-115. doi: 10.1037/1082-

    989X.9.1.93

Mõttus, R., Realo, A., Allik, J., Esko, T., Metspalu, A., & Johnson, W. (2015). Within-trait

    heterogeneity in age group differences in personality domains and facets:

    Implications for the development and coherence of personality traits. *PLoS ONE*

    10(3): e0119667. doi:10.1371/journal.pone.0119667

Osterlind, S. J., Everson, H. T. (2009). *Differential item functioning (2nd ed.).* Thousand

    Oaks, CA: Sage.

Putnick, D. L. & Bornstein, M. H. (2016). Measurement invariance conventions and

    reporting: The State of the art and future directions for psychological research.

    Developmental Review, 41, 71-90. doi:10.1016/j.dr.2016.06.004

Psychological Corporation (2002). *WAIS-III/WMS-III technical manual (updated)*. San

     Antonio, TX: Author.

R Core Team (2018). *R: A language and environment for statistical computing.* R Foundation

     for Statistical Computing, Vienna, Austria. (https://www.R-project.org/)

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of

     Statistical Software, 48*(2), 1-36. doi:10.18637/jss.v048.i02

Roth, B., Becker, N., Romeyke, S., Schäfer, S., Domnick, F., & Spinath, F. M. (2015).

     Intelligence and school grades: A meta-analysis. *Intelligence, 53*, 118-137.

     doi:10.1016/j.intell.2015.09.002

Saklofske, D. H., Hildebrand, D. K., & Gorsuch, R. L. (2000). Replication of the factor

     structure of the Wechsler Adult Intelligence Scale – Third edition with a Canadian

     sample. *Psychological Assessment, 12*(4), 436-439. doi:10.1037/1040-3590.12.4.436

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in

     cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

     doi:10.1086/209528

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial

     measurement invariance enough? *Methodology: European Journal of Research

     Methods for the Behavioral and Social Sciences, 9*(1), 1-12. doi:10.1027/1614-

     2241/a000049

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of

     longitudinal research. *Intelligence, 35*, 401-426. doi:10.1016/j.intell.2006.09.004

Zhu, J. & Chen, H.-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29*(6), 570-580. doi:10.1177/0734282910396323

Taub, G. E., McGrew, K. S., & Witta, E. L. (2004). A confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale-Third edition. *Psychological Assessment, 16*(1), 85-89. doi:10.1037/1040-3590.16.1.85

Tommasi, M., Pezzuti, L., Colom, R., Abad, F.J., Saggino, A., & Orsini, A. (2015). Increased educational level is related with higher IQ scores but lower g-variance: Evidence from the standardization of the WAIS-R for Italy. *Intelligence, 50*, 68-74. doi:10.1016/j.intell.2015.02.005

Tulsky, D. S., & Price, L. R. (2003). The joint WAIS-III and WMS-III factor structure: Development and cross-validation of a six-factor model of cognitive functioning. *Psychological Assessment, 15*, 149-162. doi:10.1037/1040-3590.15.2.149

Tran, U. S., Hofer, A. A., & Voracek, M. (2014). Sex differences in general knowledge: meta-analysis and new data on the contribution of school-related moderators among high-school students. *PloS ONE, 9*(10), e110391. doi:10.1371/journal.pone.0110391

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-70. doi:10.1177/109442810031002

Van der Maas, H. L., Dolan, C.V., Grasman, R. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*(4), 842-861. doi:10.1037/0033-295X.113.4.842

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology, 9*(4), 486-492. doi:10.1080/17405629.2012.686740

Van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J. C., Colom, R., & Boomsma, D. I. (2006). Sex differences on the Dutch WAIS-III. *Intelligence, 34*, 273-289. doi:10.1016/j.intell.2005.08.002

Ward, L. C., Ryan, J. J., & Axelrod, B. N. (2000). Confirmatory factor analyses of the WAIS-III standardization data. *Psychological Assessment, 12*(3), 341-345. doi: 10.1037//1040-3590.12.3.341

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale, 3rd ed*. San Antonio: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition*. San Antonio, TX: Pearson.

Wechsler, D. (in press). *WAIS-III läbiviimise ja skoorimise juhend – täiendatud versioon. [Estonian WAIS-III administration and scoring manual – revised edition.]* Tallinn: Tänapäev.

Wicherts, J. M. (2016). The importance of measurement invariance in neurocognitive ability testing. *The Clinical Neuropsychologist, 30*(7), 1-11. doi:10.1080/13854046.2016.1205136

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 20*(3), 39-47. doi: 10.1111/j.1745-3992.2010.00182.x

Yoon, M., & Kim, E. S. (2014). Comparison of sequential and non-sequential specification

    searches in testing factorial invariance. *Behavior Research Methods, 46*(4), 1199-

    1206. doi:10.3758/s13428-013-0430-2

Table 1

Fit Indices for Tested Confirmatory Factor Analysis Models

Model	χ^2	df	CFI	TLI	RMSEA	90% C.I. RMSEA
1. One *g* factor	1812.994	65	.752	.703	.188	.180 - .195
2. Two factors	897.277	64	.882	.856	.131	.123 - .138
3. Four factors	277.576	59	.969	.959	.070	.061 - .078
3a. Four factors – Arithmetic loading on VC factor	270.759	59	.970	.960	.068	.060 - .077
3b. Four-factor model – Arithmetic split on VC and WM factors	172.240	58	.984	.978	.051	.042 - .060
4. Hierarchical model	371.793	61	.956	.944	.082	.074 - .090
4a. Hierarchical model – Arithmetic split on VC and WM	227.062	60	.976	.969	.060	.052 - .069
5. CHC-based model	408.03	60	.951	.936	.087	.079 - .095
6. CHC-based model – Arithmetic split on Gf and Gsm factors	361.98	59	.957	.943	.082	.074 - .090

Note. VC = Verbal Comprehension; WM = Working Memory; CHC = Cattell-Horn-Carroll;
Gf = fluid reasoning factor; Gsm = short-term memory factor; df = degrees of freedom; CFI =
comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of
approximation; C.I. = confidence interval.

Table 2

Descriptive Statistics for WAIS-III Data Stratified by Sex

WAIS-III subtest/scale	Male			Female			Effect size d	t-test
	N	M	SD	N	M	SD		
Vocabulary	341	9.87	3.11	428	10.16	2.80	-0.10	-1.33
Similarities	341	10.02	2.94	429	10.07	2.89	-0.02	-0.26
Arithmetic	341	10.53	3.10	429	9.70	2.72	0.28	3.91***
Digit Span	341	9.99	3.16	429	10.01	2.62	-0.01	-0.11
Information	341	10.53	2.96	429	9.49	2.84	0.36	4.92***
Comprehension	340	10.11	2.97	428	9.98	2.90	0.04	0.61
Letter-Number S.	340	9.88	3.07	428	10.23	2.71	-0.12	-1.68
Picture Completion	341	10.18	2.90	429	9.93	3.00	0.08	1.14
Coding	340	9.39	2.87	428	10.58	2.67	-0.43	-5.95***
Block Design	341	10.15	2.92	429	9.85	2.75	0.11	1.46
Matrix Reasoning	341	10.23	3.04	429	9.89	2.76	0.12	1.61
Picture Arrangement	341	10.23	3.12	429	9.71	2.85	0.17	2.65*
Symbol Search	341	9.72	2.84	429	10.21	2.68	-0.18	-2.451*
Object Assembly	341	9.93	3.06	429	10.12	2.81	-0.06	-0.93
VCI	341	100.87	15.51	428	99.42	14.44	0.10	1.34
POI	341	101.02	15.25	429	99.11	14.38	0.13	1.79
WMI	340	100.61	16.88	428	99.45	13.40	0.08	1.04
PSI	340	97.11	15.38	428	102.08	14.12	-0.34	-4.65***
Verbal IQ	340	101.04	15.91	428	99.04	14.05	0.13	1.83
Performance IQ	340	100.32	15.54	428	99.82	14.47	0.03	0.46
Full Scale IQ	339	100.75	15.75	427	99.21	13.95	0.10	1.43

Table 2 (continued)

Note. Letter-Number S. = Letter-Number Sequencing; VCI = Verbal Comprehension Index;

POI = Perceptual Organization Index; WMI = Working Memory Index; PSI = Processing

Speed index. VCI, POI, WMI, PSI, Verbal IQ, Performance IQ and Full Scale IQ are

composite scores. Effect size d = Cohen's d. T-test coefficients reflect independent samples t-

tests between men and women. Positive t-values and effect sizes indicate male advantage;

negative t-values indicate female advantage. *$p < 0.05$; ***$p < 0.001$.

Table 3

Goodness-of-Fit Indices for Testing Measurement Invariance between Males and Females with Multi-Group Confirmatory Factor Analysis

Model	χ^2	df	RMSEA	CFI	ΔCFI
1. Configural invariance	214.27	116	.047	.986	-
2. Metric invariance	228.89	126	.046	.985	.001
3. Scalar invariance	337.72	135	.063	.971	.014
3a. Releasing Information intercept	301.405	134	.057	.976	.009
3b. Releasing Arithmetic intercept	268.38	133	.052	.981	.004
3c. Releasing Coding intercept	249.20	132	.048	.983	.002

Note. The metric model was compared to the configural mode; the scalar models 3 to 3c were compared to the metric model. CFI = comparative fit index; df = degrees of freedom; RMSEA = root mean square error of approximation.

Table 4

Male and Female Means and Standard Deviations of the Latent Factors

		VC	WM	PS	PO
Females	Mean	0	0	0	0
	SD	1	1	1	1
Males	Mean	.038	.076	.125	.369***
	SD	1.385	1.348	1.533	1.477

Note. VC = Verbal Comprehension; WM = Working Memory, PS = Processing Speed; PO = Perceptual Organization. ***$p < 0.001$.

Table 5

Descriptive Statistics for WAIS-III Data Stratified by Age Groups

WAIS-III subtest	Age group 16-29 years			Age group 30-54 years			Age group 55-89 years		
	N	*M*	*SD*	*N*	*M*	*SD*	*N*	*M*	*SD*
Vocabulary	242	10.06	2,87	252	9.98	3.17	275	10.06	2.80
Similarities	242	10.06	2,97	252	10.13	2.93	276	9.97	2.84
Arithmetic	242	10.24	2,99	252	9.97	3.10	276	10.00	2.69
Digit Span	242	10.08	2.79	252	10.05	3.07	276	9.89	2.75
Information	242	10.06	2.86	252	9.94	3.00	276	9.86	2.96
Comprehension	242	10.16	3.05	252	9.94	2.90	274	10.03	2.86
Letter-Number S.	241	10.28	2.93	252	10.14	2.76	275	9.83	2.93
Picture C.	242	10.01	2.93	252	10.23	3.08	276	9.89	2.87
Coding	241	10.09	2.92	252	10.17	2.90	275	9.91	2.65
Block Design	242	10.17	3.00	252	10.06	2.95	276	9.74	2.54
Matrix Reasoning	242	9.98	2.94	252	10.26	2.92	276	9.89	2.83
Picture A.	242	10.06	3.11	252	10.18	2.93	276	9.62	2.89
Symbol Search	242	10.04	3.00	252	10.14	10.14	276	9.80	2.53
Object Assembly	242	10.09	2.95	252	10.08	10.08	276	9.95	2.79
VCI	242	100.29	14.96	252	100.08	15.25	275	99.84	14.66
POI	242	100.32	15.09	252	101.05	15.18	276	98.64	14.10
WMI	241	100.97	15.52	252	100.01	15.50	275	99.03	14.16
PSI	241	100.18	16.00	252	100.68	15.08	275	98.87	13.65
Verbal IQ	242	100.56	15.31	252	99.81	15.32	274	99.46	14.23
Performance IQ	241	100.60	15.52	252	101.27	15.21	275	98.44	14.10
Full Scale IQ	241	100.58	15.32	252	100.38	15.24	273	98.82	13.84

Table 5 (continued)

Note. Letter-Number S. = Letter-Number Sequencing; Picture C. = Picture Completion;

Picture A. = Picture Arrangement; VCI = Verbal Comprehension Index; POI = Perceptual

Organization Index; WMI = Working Memory Index; PSI = Processing Speed Index. VCI,

POI, WMI, PSI, Verbal IQ, Performance IQ and Full Scale IQ are composite scores.

Table 6

Goodness-of-Fit Indices for Testing Measurement Invariance across Age Groups with Multi-Group Confirmatory Factor Analysis

Model	χ^2	df	RMSEA	CFI	ΔCFI
1. Configural invariance	266.91	174	.046	.984	-
2. Metric invariance	319.05	194	.050	.978	.006
2a. Releasing Block Design loading to PO	305.29	192	.048	.980	.004
2b. Releasing Matrix Reasoning loading to PO	293.64	190	.046	.982	.002
3. Scalar invariance	382.88	208	.057	.969	.013
3a. Releasing Picture Arrangement intercept	347.97	206	.052	.975	.007
3b. Releasing Arithmetic intercept	334.41	204	.050	.977	.005
3c. Releasing Vocabulary intercept	323.96	202	.049	.979	.003
3d. Releasing Information intercept	308.71	200	.046	.981	.001

Note. The metric models 2 to 2b were compared to the configural model; the scalar models 3 to 3d were compared to the metric model 2b. PO = Perceptual Organization; CFI = comparative fit index; df = degrees of freedom; RMSEA = root mean square error of approximation.

Table 7

Differences of Latent Factors Between Age Groups

		VC	WM	PS	PO
Age 16-29	Mean	0	0	0	0
	SD	1	1	1	1
Age 30-54	Mean	.107	-.333**	-.824***	-.279**
	SD	1.524	1.761	1.857	1.556
Age 55-89	Mean	-.331**	-1.407***	-2.471***	-1.705***
	SD	1.695	2.359	2.725	2.143

Note. VC = Verbal Comprehension; WM = Working Memory; PS = Processing Speed; PO = Perceptual Organization. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Table 8

Descriptive Statistics for WAIS-III data Stratified by Educational Levels

WAIS-III subtest/scale	Educational level: basic			Educational level: secondary			Educational level: higher			r_s
	N	M	SD	N	M	SD	N	M	SD	
Vocabulary	121	7.59	2.79	389	10.04	2.60	152	12.11	2.51	.499**
Similarities	122	7.67	2.79	389	10.05	2.59	152	11.95	2.45	.467**
Arithmetic	122	8.31	2.93	389	10.02	2.77	152	11.45	2.49	.336**
Digit Span	122	8.69	2.97	389	10.05	2.71	152	10.93	2.85	.263**
Information	122	7.49	2.59	389	9.93	2.72	152	11.88	2.38	.424**
Comprehension	121	7.69	2.82	388	10.07	2.61	152	11.99	2.53	.470**
Letter-Number S.	122	8.25	3.17	387	10.01	2.69	152	11.32	2.22	.342**
Picture C.	122	8.60	3.26	389	10.30	2.85	152	10.64	2.66	.156**
Coding	121	8.11	3.08	388	10.15	2.51	152	11.32	2.53	.350**
Block Design	122	8.20	2.99	389	10.21	2.73	152	10.82	2.38	.256**
Matrix Reasoning	122	8.13	3.02	389	10.12	2.66	152	11.36	2.62	.340**
Picture A.	122	8.52	3.07	389	9.98	2.90	152	10.75	2.81	.231**
Symbol Search	122	8.37	2.95	389	10.21	2.58	152	10.81	2.60	.369**
Object Assembly	122	8.88	3.37	389	10.27	2.76	152	10.49	2.87	.146**
VCI	121	86.29	12.96	389	99.87	12.91	152	111.67	12.21	.528**
POI	122	89.72	15.23	389	101.03	13.86	152	105.63	12.93	.328**
WMI	122	90.07	15.29	387	99.68	13.86	152	107.34	12.70	.360**
PSI	121	89.50	15.55	388	100.76	13.74	152	105.99	13.22	.339**

Table 8 (continued)

|										
Verbal IQ	121	86.63	13.03	388	99.67	13.00	152	111.22	12.59	.512**
Performance IQ	121	88.92	15.12	388	100.82	13.77	152	106.81	13.31	.364**
Full Scale IQ	120	86.75	13.26	387	99.98	13.05	152	109.99	12.39	.488**

Note. Letter-Number S. = Letter-Number Sequencing; Picture C. = Picture Completion; Picture A. = Picture Arrangement; VCI = Verbal Comprehension Index; POI = Perceptual Organization Index; WMI = Working Memory Index; PSI = Processing Speed Index. VCI, POI, WMI, PSI, Verbal IQ, Performance IQ and Full Scale IQ are composite scores. **$p <$ 0.01.

Table 9

Goodness-of-Fit Indexes for Testing Measurement Invariance across Educational Levels with

Multi-Group Confirmatory Factor Analysis

Model	$\chi2$	df	RMSEA	CFI	ΔCFI
1. Configural invariance	267.82	174	.050	.982	-
2. Metric invariance	322.76	194	.055	.975	.007
2a. Releasing Information loading to VC	306.11	192	.052	.978	.004

Note. The metric models were compared to the configural model. CFI = comparative fit index;

df = degrees of freedom; RMSEA = root mean square error of approximation.
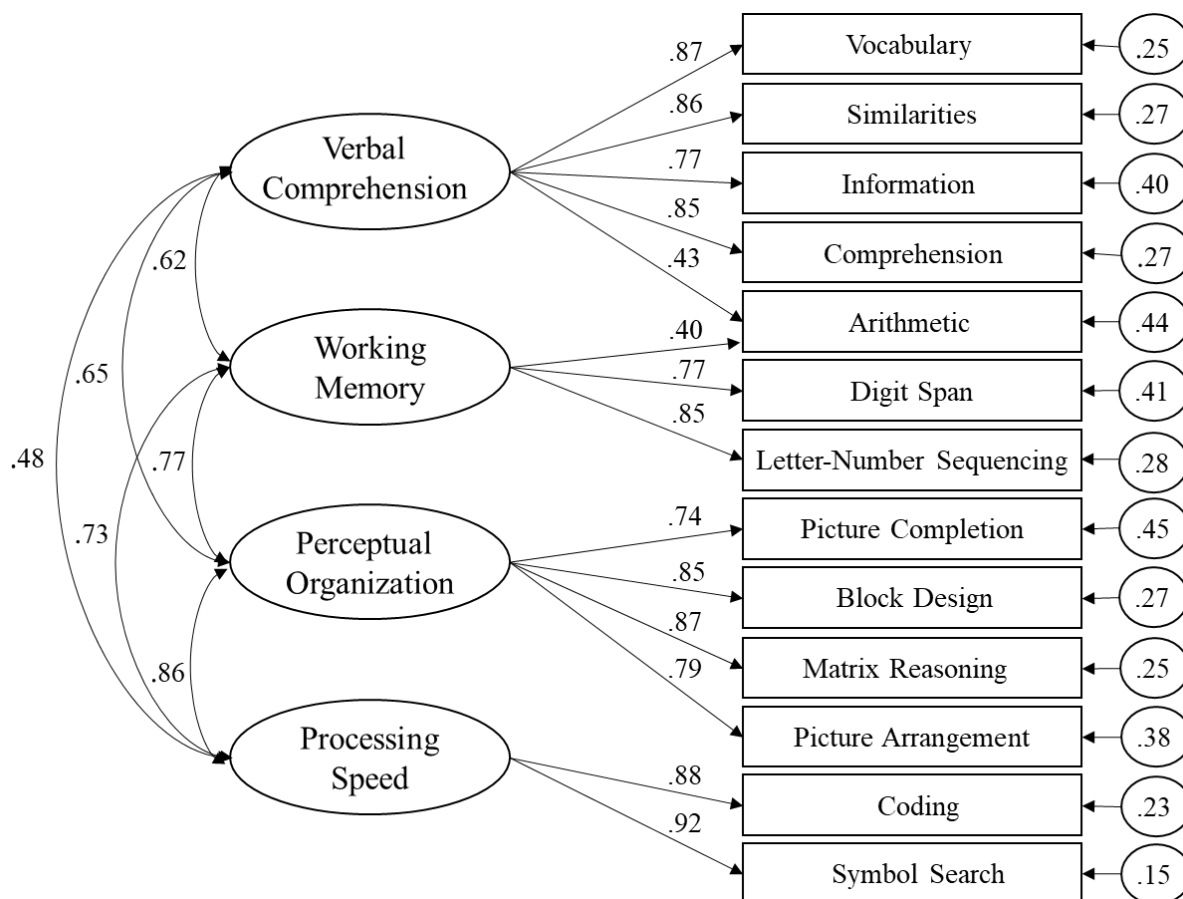
Supplemental Figure 1. Path diagram to present the standardized factor loadings and covariances between factors for the four-factor model when Arithmetic was allowed to load on both the Verbal Comprehension and the Working Memory factors. This was chosen as baseline model for measurement invariance analyses as it provided the best overall fit to the data $-\chi^2(58) = 179.36$, CFI = .984, TLI = .978, RMSEA = .051.