Testing Measurement and Structural Equivalence in Different Age Groups of Children.

Natacha Borgers[1]
University of Amsterdam

Bea van den Bergh
Centrum voor Bevolkings- en Gezinsstudie, Brussel

Joop Hox
Utrecht University

Abstract

In this study the measurement and structural equivalence was tested across two instruments, the original Dutch version of the *Competentie-Beleving Schaal voor Kinderen* (CBSK) and a shortened and simplified version of the same test across different ages. The CBSK is intended for children between 8 and 12 years old. The shortened version is developed for children aged between 6 and 8 years old. To use the shortened CBSK as a counterpart of the original CBSK, both instruments should be equivalent across ages. The analyses have been done on two data sets. The first consists of a sample of 75 nine years old children who responded on both instruments. The second data set consists of a sample of 951 children aged between six and nine years old and responded on the shortened CBSK and 758 children between nine and 13 years old who responded on the original CBSK.

Structural equation modeling was used to test the measurement and structural equivalence. Given the small sample size of one of the groups, bootstrap methods were used to assess the fit of the models.

The results show that both instruments are reasonably well comparable and exchangeable for children aged nine years old and younger. However this conclusion does not persist for responses of children aged ten years or older.

[1] Corresponding author: Natacha Borgers
Faculty of Social and Behavioral Sciences, university of Amsterdam
P.O.Box 94208, 1090 GE Amsterdam , The Netherlands
tel: +31 20 5251526; email: nborgers@educ.uva.nl

1. INTRODUCTION

Self-esteem is an influential aspect in the development of children and adolescents. Low self-esteem often underlies behavioral and emotional problems, which impedes an optimal development of children. For an overview of this field over the previous 15 years, integrating different perspectives (psycho-analysis, social psychology, cognitive psychology), see Harter (1998).

Studies of children's self-concept have been dominated by a focus on the self, studied as a uni-dimensional concept and without giving attention to developmental changes. Harter, among others, considers self-concept as a multidimensional phenomenon, consisting of several aspects (domains). The self-concept contains domain specific self-concepts and a global self-evaluation. Harter calls these domain specific self-concepts 'self-perceptions'. For example: self-perception on sports performance. Global self-worth is based on self evaluations made along different dimensions and their relevance to a global sense of the self. Which specific domains are important, depends on the age of the child. For children at primary school for example, the important specific domains are: school, sports, their appearance, their own behavior and contact with peers. With age the different domains change in content or supplementary domains appear. Regarding global self-worth, Harter states that it does not develops before the age of eight years old. Children from the age of eight and up will be conforming to the expectations of others regarding the type of person they want them to be. Children internalize these expectations as self-guides. This internalization is in turn used to evaluate the global self. Although global self-worth and domain specific self-concepts are related, the former is not the sum of the domain specific self-concepts. Global self-worth is the result of a comparison between someone's pretensions and their successes on the various domains and the possibilities to discount the importance of the various domains in the global self-worth.

Contrary to Harter (Harter, 1998), different developmental psychologists determined that children as young as five years old have a global self-worth at their disposal. Verschueren and Marcoen for example (1999) developed a special individual interview to determine 'the global-self' of young children. In addition, recent opinions about the development of social cognitions lead to the conclusion that children younger than eight years old have a global self-worth at their disposal (Flavell, 2000; Steerneman & Pelzer, 1994).

As part of a large survey 'Leefsituatie Kinderen' for children at primary education in Flanders (Van den Bergh, 1995), Veerman's et al. (1994; 1997) Dutch version of Harter's (1985) *Self-Perceptions Profile for Children* (SPPC), the *Competentie-Beleving Schaal voor Kinderen* (CBSK) is used to research the self-concept of children at the primary school age. The SPPC is a scale with good psychometric qualities, and with good cross-cultural equivalence (Byrne, 1996; Van den Bergh & Ranst, 1998). The same applies to Veerman's et al. Dutch version (Van den Bergh, 1996; Van den Bergh & Marcoen, 1999; Veerman et al., 1997). The SPPC and the CBSK are intended for the self-concept of children between 8 and 12 years old. However, the purpose of this survey was to examine the self-concept of children between 6 to 12 years old. Therefore, a self-concept scale was needed for this younger age group. However, the available self-concept scales for young children must be administered individually, under the guidance of a test assistant. For example, Harter and Pike (1980; 1984) developed the 'Pictorial Scale of Perceived Competence and Social Competence'. A test assistant administers this scale to an individual child. The test assistant shows pictures, one by one, to the child. Besides the time-consuming administration of responses, this scale measures only the domain specific self-concepts and not the global self-worth. This is in accordance

with Harter's theoretical principle that the global self-worth does not develop until the age of eight years old. Verschueren and Marcoen (1993) developed a self-concept scale for young children, which measures the global self-worth but can only be administered individually with guidance of a test assistant.

Consequently, the decision was made to simplify the CBSK for the 'leefsituatie'-survey, to make it possible to measure the self-concept of the younger children with a written questionnaire. In a Pilot-study it was determined that the wording of the questions was too abstract for these young children, the scale was too long, and the response format was too complex. Therefore, the questions were made more concrete and the scale was shortened from 36 to 18 items in total (see Van den Bergh & Ranst, 1998), and the response format was made simpler. In the shortened version the same subscales where used as in the original CBSK, including a scale to measure the global self-worth. The selection of the items is based on the 'Teacher's Rating Scale of Child's Actual Behavior (Harter, 1985). This scale measures the perception of teachers for each of the specific domains. The global self-worth subscale is based on the Dutch version of the Teacher's Rating scale, which was also developed for the 'Leefsituatie'-survey (Van den Bergh, 1999; Van den Bergh & Marcoen, 1999).

To use this shortened CBSK scale as a counterpart of the original CBSK, it is important to assess if both instruments are equivalent. Do both instruments measure the same construct, on the same scale? Accordingly, the purpose of this study was to test the factorial equivalence across both instruments. The question of factorial equivalence in this study focuses on the correspondence of factors across both instruments, and on the equivalence of measurement and structure (Byrne, Shavelson, & Muthén, 1989). Testing for equivalence of measurement can be formulated as the assumption that the items of the shortened CBSK scale are perceived in the same way as the items of the CBSK and with the same degree of accuracy. Testing for structural equivalence can be reformulated as the assumption that both instruments show the same empirical results, show the same factor mean and variance structure.

The translation of the CBSK into a more simplified version, the shortened CBSK is in fact a reformulation of a part of the items. Two rival hypotheses can be formulated about the effects of the reformulation on the relationship between variables. The first one states that, even if formulation effects exist when univariate statistics are compared, this does not necessarily imply an effect on multivariate statistics. The reasoning is that the observed differences just reflect a shift of the position on a specific variable. This line of reasoning is sometimes called 'the form-resistant correlation hypothesis' (Krosnick & Alwin, 1987; de Leeuw, 1992). The second hypothesis derives from statistical distribution theory. This hypothesis states that rather small differences in responses can cause dramatic change in multivariate statistics (de Leeuw, 1992). To compare the responses of the younger children on the shortened CBSK with the responses of the older children on the original CBSK the instruments should show at least equality of measurement, which would reject the second hypothesis.

2. METHOD

2.1 Data

The first data set we used consists of a sample of 75 children who participated in the research on the effects of stress during pregnancy of Van den Bergh (1989; 1990). The children completed the questionnaires individually, during a house visit as part of a follow-up study (Van den Bergh, 1989; Van den Bergh, 1990). The children answered the CBSK

questionnaire as well as the short version of the CBSK. The age of the children who participated in this small study was during the data collection between eight years and ten months and nine years and two months old.

The second data set used was collected as part of a large survey 'Leefsituatie Kinderen' for children at primary education in Flanders (Van den Bergh, 1995; Van den Bergh, 1997). The questionnaires were administered in a class setting. The sample contains 68 primary schools and is representative for the Flemish Community (Belgium). This study examined the competence of six- to thirteen-year-old children and the nature and quality of their living conditions. The total sample size contained 1709 children, 758 children between nine and 13 years old and 951 children between six and nine years old. Trained research assistants administered the data in class sessions in the regular classrooms of the children.

For the analyses in this study we just included those children who responded on every question of the instruments. To achieve a clear separation of the different age groups, we excluded the nine years old from the second data set. Finally, the first data set left 66 children for our analysis. The second data set ended up with 840 children between six and eight years old and 464 children between the age of ten and 13 years old.

2.2 Instruments

For this study two instruments were used. First, Veerman's et al. (1994; 1997) Dutch version of Harter's (1985) *Self-perceptions profile for Children* (SPPS), the *Competentie Beleving Schaal voor Kinderen* (CBSK). This instrument is intended for children between eight and 12 years old. The CBSK contains 36 items that can be divided into six subscales with six items. Five of these subscales measure a distinct and specific domain of the self-concept. These domains are respectively, Scholastic Competence, Social Acceptance, Athletic Competence, Physical Appearance and Behavioral Conduct. The sixth scale measures Global Self-worth. All items are formulated as bipolar statements, for example 'Some kids often forget what they learn' but 'Other kids can remember things easily'. First the child has to decide which statement fits the best. Secondly the child has to report whether the chosen statement is 'sort of true' or 'really true' for him/her. Each item is scored from 1 to 4; a high score reflects a higher degree of perceived competence.

The second instrument, the shortened CBSK scale is developed specifically for younger children. This instrument contains likewise five subscales for the specific domains of self-concept and one subscale for global self-worth. This scale is based on Harter's Teacher Rating Scale of Child's Actual Behaviour (TRS, Harter, 1985). This instrument contains five subscales with three items each. There is no scale in this instrument to measure the Global self-worth. The five subscales are parallel tot the five specific domains of self-concept in the CBSK (and SPPS). For the subscale Global self-worth, three additional items were constructed. Compared to the items in the CBSK (and SPPC), the items in the shortened CBSK scale were formulated more concretely, and the bipolar formulation of the question was abandoned in this version, because it was too complex for the younger children. An example 'Do you sometimes forget what you have learned?' with the following response options 'no never', 'sometimes', 'often' and 'very often'. Every item is just like the CBSK scored from 1 to 4 and a high score reflects a high degree of perceived competence. To make answering the questions easier, the response categories were visualized as rectangles of increasing size, with the smallest (indicated with just a dot) representing "not at all" and the largest representing "very much".

2.3 Analysis

In the first part of this article, all analyses have been done on the first data set in which the children responded on both instruments. First the difference in reliability (Cronbach's alpha) of both instruments is tested (Feldt, 1959; Hakstian & Whalen, 1976) by means of the program Alfatest (Hox, 1991). Furthermore, the reliability of the CBSK is calculated, with Spearman Brown formula for test extension (or shortening), as if there were only three items in the scale as it is for the shortened CBSK scale. This calculated alpha could be compared directly to the alpha of the shortened scale. The same is done for the shortened CBSK scale; the reliability was calculated for the scale as if it contained six items and compared to the reliability of the CBSK.

The aim of this study is to test the factorial equivalence (Byrne et al., 1989) between the 6-item CBSK and the shortened 3-item CBSK. To test the factorial equivalence between both instruments we performed six confirmatory factor analyses using AMOS (Arbuckle & Wothke, 1999). Testing for factorial equivalence is a twofold procedure: a) testing for equivalence of measurement and b) testing for equivalence of structure (Byrne et al., 1989). In this study, the analysis was done in two steps. First, we focused on the equivalence of measurement. For this part, the base model was the same in all six analyses, a confirmatory two-factor model in which both factors represent the corresponding subscales. In this model the factor structure is identical for both instruments. The base factor model is presented in Figure 1. A covariance (correlation) between both factors is permitted, as is the covariance (correlation) between the error terms of the corresponding items of both instruments. The factor loadings of the error terms have been fixed to one. For three out of the six subscales it was necessary to identify the model by setting the variances of both factors equal to 1 (as opposed to constraining one of the loadings). This restriction was necessary because the correlation between both factors was estimated to be higher than one (for a discussion of identification problems see Bollen, 1989). There are no equality constraints between both instruments. If this model has a good overall fit, it can be concluded that both instruments are congeneric tests.

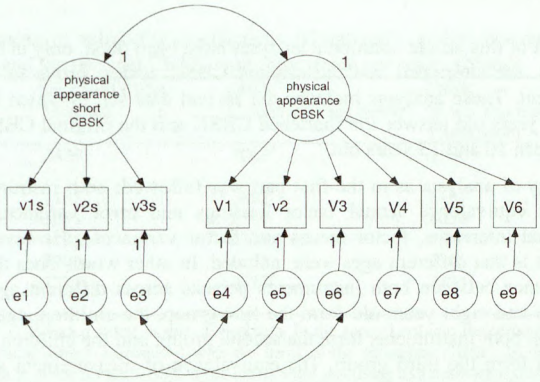

Figure 1: Confirmatory factor model for testing factorial equivalence. An example for the subscale Physical Appearance

Next, both instruments were tested for tau-equivalence. In the tau-equivalent model, the factor loadings were restricted to be equal across the corresponding items. The third model tests for parallel tests. In the parallel test model, additional equality restrictions have been imposed on the error variances. The error variances of the corresponding items of both instruments are restricted to be equal.

Secondly, the instruments were tested for equivalence of structure. The structural model in the analysis addresses the equivalence of factor means and factor variance-covariance structures (Byrne et al., 1989). In this study this is tested as follows. First we tested for equality of the factor means and secondly for the equality of the factor variances. The results of testing for equivalence of measurement determined the model that was used as base model for the second part of testing for factorial equivalence. This means that the different subscales could have different base models in this part of our study. The minimum restriction to compare factor means was the equality constraints on the factor loadings between both instruments and equality constraints on the intercepts of the observed variables across both instruments (Hox & Bechger, 1998). If the model with equal factor loadings across the instruments was rejected in the first part of this study comparing factor means did not make sense and, these tests were not performed. Besides, if the model with equality constraints on the intercepts was rejected, testing the comparability of factor means was not performed either.

Using structural equation models with small sample sizes and possibly non-normal data, as we have to, causes some problems. For small samples and non-normal data the χ^2 estimate for goodness of fit does not behave very well. In general the right tail of the empirical sampling distribution is too heavy and the χ^2 will be too high. As a consequence the specified model is rejected too often (Bollen & Stine, 1993; Boomsma, 1983). To assess the overall model fit we used the Bollen-Stine bootstrap (Bollen & Long, 1993; Bollen & Stine, 1993; Stine, 1989; Yung & Chan, 1999) for the overall fit of the model (χ^2; and bootstrapped p-value). All other fit indices are based on the asymptotic, biased χ^2, so we did not use one of these indices. Hartman, Hox, Erol, Mellenbergh, Oosterlaan, Shalev, Auerbach, Fonseca, Nøvik, Roussos, Zilber, and Sergeant (1999) show that the other bootstrapped fit indices show the same pattern as the bootstrapped chi-square and its p-value.

In the second part of this article additional analyses have been done, only in those cases where both instruments, the shortened and the original CBSK scales, prove to be equivalent or partially equivalent. These analyses included the second data set, in which the children aged between 6 and 8 years old answer the shortened CBSK and the original CBSK was answered by children between 10 and 13 years old.

The same strategy of analysis as in the first part was followed; both instruments were tested on measurement equivalence (equal factor loadings and error variances) and structural equivalence (equal intercepts, factor means and factor variances). However, the difference with the first part is that different ages were included. In other words does the equivalence or partially equivalence between both instruments persists across different ages? The children aged between six and eight years old form the first group, the children, aged nine years old, who responded on both instruments form the second group, and the children aged between 10 and 13 years old form the third group. The equivalence of measurement and structure was tested between the three groups. Figure 2 shows the visualized base model, which is used for these analyses.

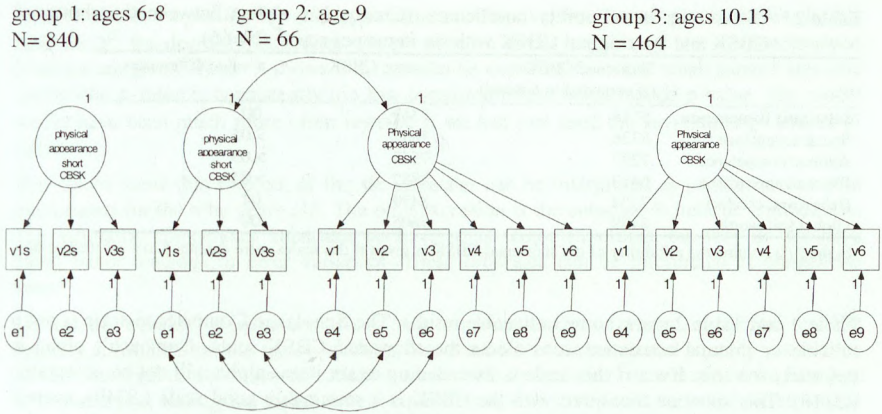

Figure 2: Testing measurement equivalence between the three groups (different ages). An example for the subscale Physical Appearance

3. RESULTS

The reliability of the six subscales in both instruments was compared in two ways. First we compared the reliability of the shortened CBSK scale as measured in our data set with the reliability of the 3 corresponding items of the CBSK and the reliability of the shortened (to three items with the Spearman Brown test extension formula) CBSK. The results are presented in Table 1. Table 2 shows the results of the second comparison. In this table the reliabilities of the extended shortened CBSK scale (extended to six items) with the CBSK as measured in our data set are compared.

Table 1 Differences in reliability coefficients (Cronbach's alpha) between the shortened CBSK and the original CBSK with three items per scale* (N=66)

	Shortened CBSK (3 items)	CBSK (3 corresponding items)	CBSK (if test shortened to 3 item scale)	p-value differences
Scholastic competence	.3439[1]	.7062[2]	.7166[3]	1-2: p =.01 1-3: p =.01
Social acceptance	.7132	.5892	.6197	p > .05
Athletic competence	.5633	.6767	.6058	p > .05
Physical appearance	.4785	.5721	.5850	p > .05
Behavioral conduct	.5790	.6333	.7360	p > .05
Global self-worth	.7375[1]	.4770[2]	.3894[3]	1-2: p =.02 1-3: p =.01

* The original CBSK is included twice in this comparison. In the second column the reliabilities of the scales is given for the three items which correspond with the items in the shortened CBSK. In the third column, the reliability of the scales is given for the CBSK if the scales were shortened to a three-item scale.

Table 2 Differences in reliability coefficients (Cronbach's alpha) between the shortened CBSK and the original CBSK with six items per scale* (N=66)

	Shortened CBSK (if extended to 6 items)	6 items CBSK	p-value differences
Scholastic competence	.5118	.8349	.00
Social acceptance	.8326	.7652	>.05
Athletic competence	.7207	.7545	>.05
Physical appearance	.6473	.7382	>.05
Behavioral conduct	.7334	.8479	.04
Global self-worth	.8489	.5605	.00

*In the first column the reliability of the shortened CBSK is given if the scales were extended to a six item scale.

Table 1 and Table 2 show some noticeable results. The Scholastic Competence-scale is not a reliable or internal consistent subscale in the shortened CBSK scale. Cronbach's alpha is definitely too low. Even if this scale is extended up to six items alpha will not be acceptable (.5118). This subscale measured with the CBSK is a reasonably good scale (.8349), even if this scale is shortened to three items (.7166). The lower reliability of all other subscales in the shortened CBSK scale seems to be the result of the smaller number of items. After increasing the number of items up to six, coefficient alpha becomes reasonable to good, although an alpha of .6473 for the physical appearance-scale can be questionable. The Global Self-worth scale does not show an acceptable result for the CBSK, in our data. This is striking, because this scale usually shows a reliability coefficient above .70, like the other subscales.

Table 3 Testing for equivalence of measurement for the 6 subscales for children aged nine. N = 66*

	χ_{df}^2 ; p-value$_{assymptotiic\ \chi 2}$ (p_a); p-value$_{bootstrap}$ (p_b);											
Models	Scholastic competence[2]		Social acceptance[3]		Athletic competence[4]		Physical appearance		Behavioral conduct		Global self-worth	
Congeneric tests: equal factor structure	χ_{24}^2 =	**43.9**	χ_{24}^2 =	32.2	χ_{24}^2 =	32.5	χ_{23}^2 =	45.4	χ_{23}^2 =	24.9	χ_{23}^2 =	26.7
	p_a=	**.01**	p_a=	.12	p_a=	.12	p_a=	.00	p_a=	.36	p_a=	.27
	p_b =	**.12**	p_b =	.37	p_b =	.34	p_b =	.13	p_b =	.56	p_b =	.52
Tau-equivalent tests: + equal factor loadings	χ_{27}^2 =	86.3	χ_{27}^2 =	32.8	χ_{27}^2 =	40.7	χ_{26}^2 =	48.7	χ_{26}^2 =	32.4	χ_{26}^2 =	32.8
	p_a=	.00	p_a =	.20	p_a=	.04	p_a=	.00	p_a=	.18	p_a=	.17
	p_b =	.00	p_b =	.48	p_b =	.24	p_b =	.15	p_b =	.48	p_b =	.44
Parallel tests: + equal error variances	χ_{30}^2 =	90	χ_{30}^2 =	**36.5**	χ_{30}^2 =	**50.3**	χ_{29}^2 =	**54.2**	χ_{29}^2 =	**54.3**	χ_{29}^2 =	**48.7**
	p_a =	.00	p_a=	**.19**	p_a =	**.01**	p_a=	**.00**	p_a=	**.00**	p_a=	**.01**
	p_b =	.00	p_b =	**.46**	p_b =	**.14**	p_b =	**.16**	p_b =	**.11**	p_b =	**.18**

* The accepted model is printed bold.
The number of degrees of freedom differ slightly across the models because identification problems sometimes necessitated and additional constraint.

[2] Correlation between both factors set equal to 1 to attain identification of the model.
[3] Correlation between both factors is set equal to 1 to attain identification of the model.
[4] Correlation between both factors is set equal to 1 to attain identification of the model.

Table 3 shows the results for testing measurement equivalence for all six subscales. In this table the χ^2, the degrees of freedom, the p-values for the asymptotic χ^2 and the bootstrapped p-values are given. Table 3 shows that, as is to be expected from the small sample size, the asymptotic p-value is consistently too low compared to the bootstrapped p-value. The model would have been much more often rejected if we had just used the asymptotic χ^2 and its p-value.

The results show that five out of the six subscales can be interpreted as parallel across both instruments for the nine years old. The only exception is the subscale Scholastic competence. The scholastic competence subscales do have equal factor structures, but they do not have equal factor loadings or error variances. These subscales can be interpreted as congeneric tests.[5]

To test the equality of structure across both instruments the minimum restriction was the equality constraints on the factor loadings between both instruments and equality constraints on the intercepts of the observed variables across both instruments. As a result of the first part of our analysis, testing for measurement equivalence, five scales will be tested for structural equivalence. The results of these analyses are shown in Table 4.

The base model for these analyses is the same for all five subscales: equal factor loadings and equal error variances across both instruments. The first model tested the minimum restriction of equality constraints on the intercepts of the observed variables across both instruments. The overall model fit (bootstrapped) is given in Table 4. The next model was testing for equality of factor means. Subsequently, the variances were tested to be equal across both instruments.

Table 4 Testing for equality of structure for five out of the six subscales for children aged nine years old N = 66*.

	χ_{df}^2 ; p-value$_{assymptotiic\ \chi 2}$ (p_a); p-value$_{bootstrap}$ (p_b);				
	Social acceptance	Athletic competence	Physical appearance	Behavioral conduct	Global Self-worth
equal intercepts	$\chi_{31}^2 = 44.5$ $p_a = .06$ $p_b = .28$	$\chi_{31}^2 = 88.1$ $p_a = .00$ $p_b = .01$	$\chi_{30}^2 = 58.2$ $p_a = .00$ $p_b = .13$	$\chi_{30}^2 = 49.6$ $p_a = .01$ $p_b = .17$	$\chi_{30}^2 = 57.6$ $p_a = .00$ $p_b = .11$
equal factor means	$\chi_{32}^2 = 50.6$ $p_a = .02$ $p_b = .19$		$\chi_{31}^2 = 58.8$ $p_a = .00$ $p_b = .14$	$\chi_{31}^2 = 53.6$ $p_a = .01$ $p_b = .14$	$\chi_{31}^2 = 64$ $p_a = .00$ $p_b = .06$
equal factor variances	$\chi_{33}^2 = \mathbf{51.3}$ $\mathbf{p_a = .02}$ $\mathbf{p_b = .19}$		$\chi_{32}^2 = \mathbf{62}$ $\mathbf{p_a = .00}$ $\mathbf{p_b = .11}$	$\chi_{32}^2 = \mathbf{63.7}$ $\mathbf{p_a = .00}$ $\mathbf{p_b = .08}$	$\chi_{32}^2 = \mathbf{64}$ $\mathbf{p_a = .00}$ $\mathbf{p_b = .07}$

* The accepted model is printed bold.
The number of degrees of freedom differ slightly across the models because identification problems sometimes necessitated and additional constraint.

As can be seen in Table 4, the equality restriction on the intercepts of the observed variables for the Athletic competence-scale is rejected. The overall model fit is not acceptable. As a result is does not make sense to test equality of factor means for these scale. For the four other subscales Social acceptance, Physical appearance, Behavioral conduct and Global self-worth

[5] Note that this does not rule out the possibility of significantly different reliabilities, due to different latent factor variances.

however, the overall model fit is still reasonable with all structural equality constraints.

In the second part of this article we present the results of testing measurement and structure equivalence between both instruments across the three age groups. In Table 5 the results of the analyses testing measurement equivalent between the three groups are presented.

The results in this Table show that no scale can be conceived as tau-equivalent or parallel tests across the age groups. Besides, the Table shows that four out of the six subscales do not even have equal factor structures. The subscales scholastic competence and behavioral conduct can be interpreted as congeneric tests. However, the minimum restriction for testing structural equivalence, equality of factor loadings, is not satisfied in any of the subscales. This means that comparing factor means did not make sense and these tests were not performed.

Table 5 Testing for equivalence of measurement for the 6 subscales for three different age groups. (3 groups: Group 1: eight years and younger, N = 840; Group 2: nine years old, N = 66; Group 3: ten years old and older, N= 464)*

	χ^2_{df} ; p-value$_{assymptotiic\ \chi 2}$ (p_a); p-value$_{bootstraped\ \chi 2}$ (p_b);											
Models	Scholastic competence[6]		Social acceptance[7]		Athletic competence[8]		Physical appearance		Behavioral conduct		Global self-worth	
Congeneric tests: equal factor structure	χ^2_{33} =	**61.3**	χ^2_{33} =	85.3	χ^2_{33} =	69.3	χ^2_{32} =	70.6	χ^2_{32} =	**42.9**	χ^2_{32} =	71.1
	p_a=	**.00**	p_a=	.00	p_a=	.00	p_a=	.00	p_a=	**.09**	p_a=	.00
	p_b =	**.08**	p_b =	.00	p_b =	.03	p_b =	.04	p_b =	**.36**	p_b =	.02
Tau-equivalent tests: + equal factor loadings	χ^2_{42} =	157.1	χ^2_{42} =	94.0	χ^2_{42} =	130.0	χ^2_{41} =	116.7	χ^2_{41} =	97.2	χ^2_{41} =	97.0
	p_a=	.00	p_a=	.00	p_a=	.00	p_a=	.00	p_a=	.00	p_a=	.00
	p_b =	.00	p_b =	.01	p_b =	.00	p_b =	.01	p_b =	.00	p_b =	.01
Parallel tests: + equal error variances	χ^2_{51} =	187.7	χ^2_{51} =	135.2	χ^2_{51} =	258.3	χ^2_{50} =	234.4	χ^2_{50} =	170.8	χ^2_{50} =	204.8
	p_a=	.00	p_a=	.00	p_a=	.00	p_a=	.00	p_a=	.00	p_a=	.00
	p_b =	.00	p_b =	.00	p_b =	.00	p_b =	.00	p_b =	.00	p_b =	.00

* The accepted model is printed bold.
The number of degrees of freedom differ slightly across the models because identification problems sometimes necessitated and additional constraint.

In a study of Van den Bergh and Van Ranst (1998) the equivalence of self-concept was tested between different grades. Their study showed that the underlying structure was not equivalent for children of the 4th grade, 5th grade and 6th grade. They found real differences in the level of self-concept. However they found equivalence of structure between the 4th and 5th graders and between the 5th and 6th graders. For that reason we did the same analyses as above with children aged seven to eight years old, the nine years old and children aged ten years old. In other words we excluded the youngest children (six years old) and we excluded the children aged 11 to 13 years old from the analyses. However the same results as shown in Table 5 were found.

Because these equality test procedures were based on three groups and almost all equality restrictions were rejected, additional analyses were done. These analyses test the equivalence

[6] Correlation between both factors set equal to 1 to attain identification of the model.
[7] Correlation between both factors is set equal to 1 to attain identification of the model.
[8] Correlation between both factors is set equal to 1 to attain identification of the model.

of two out of the three age groups between both instruments. In Table 6 the results of testing equivalence between both instruments across the age groups nine years old and ten to thirteen years old are shown. These results show that two out of the six subscales can now be interpreted as tau-equivalent, the physical appearance scale and the behavioral conduct scale and one scale, the scholastic competence scale, as congeneric test. For the three other subscales all equality restriction hypothesis are rejected. The same analyses have been done for children aged ten years old, and the same results were found.

Table 6 Testing for equivalence of measurement for the 6 subscales for different age groups. (Group 1: nine years old, N = 66; Group 2: 10 – 13 years old, N= 464)*

χ^2_{df} ; p-value$_{assymptotic\ \chi 2}$ (p_a); p-value$_{bootstrap}$ (p_b);

Models	Scholastic competence[9]	Social acceptance[10]	Athletic competence[11]	Physical appearance	Behavioral conduct	Global self-worth
Congeneric tests: equal factor structure	χ^2_{33} = **61.2** p_a= **.00** p_b = **.07**	χ^2_{33} = 85.2 p_a= .00 p_b = .01	χ^2_{33} = 69.2 p_a= .00 p_b = .03	χ^2_{32} = 70.4 p_a= .00 p_b = .05	χ^2_{32} = 42.9 p_a= .09 p_b = .38	χ^2_{32} = 71.0 p_a= .00 p_b = .02
Tau-equivalent tests: + equal factor loadings	χ^2_{39} = 121.2 p_a= .00 p_b = .00	χ^2_{39} = 92.9 p_a= .00 p_b = .01	χ^2_{39} = 85.1 p_a= .00 p_b = .01	χ^2_{38} = **80.9** p_a= **.00** p_b = **.06**	χ^2_{38} = **59.1** p_a= **.02** p_b = **.19**	χ^2_{38} = 84.8 p_a= .00 p_b = .02
Parallel tests: + equal error variances	χ^2_{39} = 123.8 p_a= .00 p_b = .00	χ^2_{45} = 101.3 p_a= .00 p_b = .01	χ^2_{45} = 96.4 p_a= .00 p_b = .02	χ^2_{44} = 119 p_a= .00 p_b = .01	χ^2_{44} = 89.4 p_a= .00 p_b = .04	χ^2_{42} = 136.2 p_a= .00 p_b = .00

* The accepted model is printed bold.
The number of degrees of freedom differ slightly across the models because identification problems sometimes necessitated and additional constraint.

Table 7 shows the results of testing equivalence between both instruments across the age groups nine years old and six to eight years old. For this younger group, the equivalence hypotheses can be confirmed. These results show resemblance to the results for testing measurement equivalence between both instruments for only the nine years old. As also is shown in Table 3, the only scale that does not show equivalence in factor loadings is the subscale scholastic competence. Two out of the five subscales, can be interpreted as tau-equivalent, athletic competence and physical appearance scale. The three remainder subscales, the social acceptance, the behavioral conduct and the global self-worth scale can even be interpreted as parallel tests.

The results in Table 6 and Table 7 show some arguments to explain why we rejected the hypotheses of measurement equivalence between all three age groups. The responses of the youngest group (six to eight years old) on the shortened CBSK are comparable to the responses of the nine years old on both instruments, with exception of the scholastic competence scale. However, this comparability does not continue for the oldest group. The responses of the children between 10 to 13 years old on the CBSK cannot be compared to the responses of the nine years old on both instruments, except for the behavioral conduct and

[9] Correlation between both factors set equal to 1 to attain identification of the model.
[10] Correlation between both factors is set equal to 1 to attain identification of the model.
[11] Correlation between both factors is set equal to 1 to attain identification of the model.

physical appearance scale. For that reason the structural equivalence analyses have been done for the youngest age group and the nine years old. These results are shown in Table 8.

Table 7 Testing for equivalence of measurement for the six subscales for different age groups. (Group 1: nine years old, N = 66; Group 2: eight years old and younger, N= 840)*

χ^2_{df} ; p-value$_{assymptotiic\ \chi 2}$ (p_a); p-value$_{bootstrap}$ (p_b);

Models	Scholastic competence[12]	Social acceptance[13]	Athletic competence[14]	Physical appearance	Behavioral conduct	Global self-worth
Congeneric tests: equal factor structure	χ^2_{24}= **44.5** p_a= **.01** p_b= **.13**	χ^2_{24}= 32.7 p_a= .11 p_b= .32	χ^2_{24}= 32.9 p_a= .11 p_b= .35	χ^2_{23} = 46.0 p_a= .00 p_b= .16	χ^2_{23} = 25.2 p_a= .59 p_b= .34	χ^2_{23} = 27.1 p_a= .25 p_b= .54
Tau-equivalent tests: + equal factor loadings	χ^2_{30} = 91.3 p_a= .00 p_b= .00	χ^2_{30} = 35.7 p_a= .22 p_b= .47	χ^2_{30} = **47.2** p_a= **.02** p_b= **.18**	χ^2_{29} = **63.0** p_a= **.00** p_b= **.07**	χ^2_{29} = 39.6 p_a= .09 p_b= .35	χ^2_{29} = 45.0 p_a= .03 p_b= .26
Parallel tests: + equal error variances	χ^2_{36} = 103.5 p_a= .00 p_b= .00	χ^2_{36} = **61.4** p_a= **.01** p_b= **.10**	χ^2_{36} = 78.2 p_a= .00 p_b= .01	χ^2_{35} = 128.8 p_a= .00 p_b= .00	χ^2_{35} = **69.7** p_a= **.00** p_b= **.07**	χ^2_{35} = **76.2** p_a= **.00** p_b= **.05**

* The accepted model is printed bold.
The number of degrees of freedom differ slightly across the models because identification problems sometimes necessitated and additional constraint.

As can be seen in Table 8 the equality restrictions on the intercepts of the observed variables for the subscales athletic competence and physical appearance are rejected.

Table 8 Testing for equality of structure for 5 out of the 6 subscales for children across age groups (Group 1: 8 years and younger, N = 840; Group 2: 9 years old, N = 66)*

χ^2_{df} ; p-value$_{assymptotiic\ \chi 2}$ (p_a); p-value$_{bootstrap}$ (p_b);

Models	Social acceptance[15]	Athletic competence[16]	Physical appearance	Behavioral conduct	Global Self-worth
equal intercepts	χ^2_{39} = 68.4 p_a= .00 p_b= .07	χ^2_{33} = 77.6 p_a= .00 p_b= .01	χ^2_{31} = 95.9 p_a= .00 p_b= .00	χ^2_{37} = 59.1 p_a= .01 p_b= .20	χ^2_{37} = 72.5 p_a= .00 p_b= .08
equal factor means	χ^2_{41} = 72.4 p_a= .00 p_b= .07			not identified	χ^2_{39} = **81.3** p_a= **.00** p_b= **.05**
equal factor variances	χ^2_{42} = **72.7** p_a= **.00** p_b= **.08**				χ^2_{41} = 95.4 p_a= .00 p_b= .02

* The accepted model is printed bold. The number of degrees of freedom differ slightly across the models because identification problems sometimes necessitated an additional constraint.

As a result, it does not make sense to test the equality of mean structure for these subscales.

[12] Correlation between both factors set equal to 1 to attain identification of the model.
[13] Correlation between both factors is set equal to 1 to attain identification of the model.
[14] Correlation between both factors is set equal to 1 to attain identification of the model.
[15] Correlation between both factors is set equal to 1 to attain identification of the model.
[16] Correlation between both factors is set equal to 1 to attain identification of the model.

The behavioral conduct scale does show equivalent intercepts but a model with equal factor means between both instruments and both age groups was not identified. The global self-worth scale does show equal factor means but the model with equal factor variances is rejected. The social acceptance scale is the only scale where the overall model fit is still acceptable with all equality restrictions.

4. CONCLUSION

In this study the measurement and structural equivalence was tested between two instruments, the original Dutch version of the CBSK and a shortened CBSK for children aged nine years old. Besides these equivalences have been tested across different ages groups.

In general we can state that both instruments are reasonably well comparable and exchangeable for children aged nine years old and younger. However, this conclusion does not hold for responses of children aged ten to 13 years old. The responses of the older children on the CBSK are barely comparable with responses of children aged nine on both instruments. The responses of children aged six to eight years on the shortened CBSK to the responses of the nine years old on both instruments are comparable. However, there is one exception, the subscale Scholastic Competence. This subscale does show equal factor structures across both instruments, and across both age groups, yet the concepts are hardly comparable, and both scales measure partially something different. Besides, the scholastic competence scale is not a reliable scale in the shortened CBSK, not even if the scale would be increased up to six items, Cronbach's alpha would still be unacceptable ($\alpha = .5$).

In two ways the above-mentioned results are interesting. First, if we administer the shortened CBSK with children aged six to eight years old their responses can be compared with the responses of the nine year old children on the original CBSK. For the subscales social acceptance and global self-worth the responses can be directly compared, keeping in mind that the shortened version of the CBSK consists of three items versus the six-item CBSK. These subscales indicate measurement equality across both instruments and equal factor means. The subscale social acceptance even shows equal factor variances. For the other three subscales (athletic competence, physical appearance and behavioral conduct) the responses cannot be directly compared. Both instruments indicate different zero-points; they do not show equal factor means. Nevertheless, because these scales indicate equality of measurement the responses on the original CBSK can be equated to the responses on the shortened CBSK. This can be done with a simple regression equation. For example, for the subscale athletic competence, the regression equation to equate both tests is:

$$Y_{\text{CBSK}} = 6.91\,(1.49) + 1.49\,(.19) * X_{\text{shortened CBSK}}$$

Secondly, the cut-off point for comparability of the constructs seems to be at the age of ten. In a study of Van den Bergh & Ranst (1998) the equivalence of measurement and structure of the original CBSK across ages has been tested. The results of this study show that the responses on the subscales of children in the 4th and 5th grade (about nine and ten years old) show equivalence of measurement and structure, the same results were found for responses of children in the 5th and 6th grade (about ten and 11 years old). However, responses of the three groups were not comparable. These results are partially in accordance with Harter's theory about the self-concept. Harter states that the specific self-concept domains, that are important depends on the age of children. With age the different domains change in content or other

domains appear. The results in this study, which are also in accordance with the results of Van den Bergh & Ranst (1998), indicate that the self-concept changes after the age of ten. The different self-concept domains of children aged ten and up are not comparable with the self-concepts of the younger children. Consequently, Van den Bergh & Ranst (1998) state, when studying the evaluation of the child's self-concept from 4th to 6th grade (from nine to 11 years old) one has to take into account that not the same self-concepts are measured for these children. Therefore they suggest that children should be situated in their own age group concerning studying the self-concept.

The results in this study showed that it is valuable and feasible to translate items of an existing scale into more simplified formulations. Translating item does not result by definition in a different concept. In this study, five out of the six subscales were translated into a more simplified formulation but still measuring the same concept. This is in accordance with the so called 'the form-resistant correlation hypothesis'. The differences in formulation just cause a shift of the position on the scale score, which is also translatable. However, if a concept itself differ with age it is not the instrument that causes inequalities but the concept.

REFERENCES

Arbuckle, J. L., & Wothke, W. (1999). *Amos (Version 4.01)*. Chicago: SmallWaters corporation.

Bollen, K.A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, California: Sage.

Bollen, K. A., & Stine, R. A. (1993). Bootstrapping Goodness-of-Fit measures in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models*. Newbury, California: Sage.

Boomsma, A. (1983). *On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality*. Unpublished PhD., Rijksuniversiteit te groningen, Groningen.

Byrne, B. M. (1996). *Measuring self-concept across the life span: Issues and Instrumentation*. Washington, DC: American Psychological Association.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

De Leeuw, E. D. (1992). *Data quality in mail, telephone and face-to-face surveys*. Unpublished Ph.D., Vrije Universiteit, Amsterdam.

Feldt, L. S. (1959). A test of hypothesis that Cronbach's alpha of Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika, 34*, 363-373.

Flavell, J. H. (2000). Development of children's knowledge about the mental world. *International Journal of Behavioral Development, 24*(1), 15-23.

Hakstian, A. R., & Whalen, T. E. (1976). A K-sample significance test for independent Alpha coefficients. *Psychometrika, 41*(2), 219-231.

Harter, S. (1985). *Manual for the Self-Perception Profile for Children.* Denver, CO: University of Denver.

Harter, S. (1998). The development of Self-representations. In N. Eisenberg (Ed.), *Handbook of Child Psychology.* (Fifth ed., Vol. 2, pp. 553-617). New York: Wiley.

Harter, S., & Pike, R. (1980). *The Pictorial Scale of Perceived Competence and Social Acceptance for Young Children.* (Unpublished coding system.). Denver: University of Denver.

Harter, S., & Pike, R. (1984). The Pictorial Scale of Perceived Competence and Social Acceptance for Young Children. *Child Development, 55*, 1969-1983.

Hartman, C. A., Hox, J. J., Erol, N., Mellenbergh, G. J., Oosterlaan, J., Shalev, R. S., Auerbach, J., Fonseca, A. C., Nøvik, T. S., Roussos, A. C., Zilber, N., & Sergeant, J. A. (1999). Syndrome dimensions of the child behavior checklist and the teacher report form: A critical empirical evaluation. *Journal of Child Psychology and Psychiatry, 40*(7), 1095-1116.

Hox, J. J. (1991). *Alfatest.* Computer program. Available from http://www.fss.uu.nl/ms/jh.

Hox, J. J., & Bechger, T. M. (1998). An introduction to structural equation modeling. *Family Science Review, 11*, 354-373.

Krosnick, J., & Alwin, D. (1987). An evaluation of cognitive theory of response-order effects in survey measurement. *Public Opinion Quarterly, 51(2)*, 201-219.

Steerneman, P., & Pelzer, H. (1994). Sociale cognitie en sociale competentie bij kinderen en jeugdigen: van theorie naar praktijk. Leuven: Garant.

Stine, R. A. (1989). An introduction to bootstrap methods. Examples and ideas. *Sociological Methods and Research, 18*(2 & 3), 243-291.

Van den Bergh, B. (1989). De emotionele toestand van de (zwangere) vrouw, obstetrische complicaties en het gedrag en de ontwikkeling van de foetus en het kind tot de leeftijd van zeven maanden. Unpublished Ph.D., Katholieke Universiteit Leuven, Leuven.

Van den Bergh, B. (1990). The influence of maternal emotions during pregnancy on fetal and neonatal behavior. *Pre- and Perinatal Psychology Journal, 5*, 119-130.

Van den Bergh, B. (1995). Onderzoek 'De leefsituatie van Kinderen op schoolleeftijd'. Beknopte beschrijving van het onderzoek + vragenlijsten + bijlagen. Brussel: Centrum voor Bevolkings- en Gezinsstudie.

Van den Bergh, B. (1996). De Competentie belevingsschaal voor Kinderen: gegevens voor Vlaanderen op basis van het CBGS-onderzoek Leefsituatie van Kinderen + afzonderlijke Bijlage met Normentabellen. (Unpublished report 1996/1). Leuven: Centrum voor Ontwikkelings psychologie, Katholieke Universiteit Leuven.

Van den Bergh, B. (1997). Kindertijd. Kinderen en ouders over de leefsituatie van lagere-schoolkinderen in Vlaanderen. (CBGS Document 1997/1). Leuven: CBGS.

Van den Bergh, B. (1999). Jongens versus meisjes: zelf- en leerkrachtbeoordeling op de CBSK en CBSL. *Kind en Adolescent, 20*(2), 93-103.

Van den Bergh, B., & Marcoen, A. (1999). Harter's self-perception profile for children. Factor structures, reliability, and convergent validity in a Dutch-speaking Belgian sample of fourth, fifth and sixth graders. *Psychologica Belgica, 39*(1), 29-47.

Van den Bergh, B., & Van Ranst, N. (1998). Self-concept in children: Equivalence of measurement and structure across gender and grade of Harter's Self-Perception Profile for Children. *Journal of Personality Assessment, 70*(3), 564-582.

Veerman, J. W., Straathof, M. A. E., & Treffers, P. D. A. (1994). *Handleiding Competentiebelevingsschaal voor Kinderen CBSK.* (Internal report). Duivendrecht: Paedologisch Instituut Duivendrecht.

Veerman, J. W., Straathof, M. A. E., Treffers, P. D. A., Van den Bergh, B., & ten Brink, L. T. (1997). *Handleiding Competentiebelevingsschaal voor kinderen CBSK.* Lisse, the Netherlands: Swets & Zeitlinger.

Verschueren, K., & Marcoen, A. (1993). *De zelfbelevingsschaal voor jonge kinderen.* (Unpublished manual). Leuven: Centrum voor Ontwikkelingspsychologie, KU Leuven.

Verschueren, K., & Marcoen, A. (1999). Representation of self and socioemotional competence in kindergartners: Differential and combined effects of attachment to mother and father. *Child Development, 70*, 183-201.

Yung, Y.-F., & Chan, W. (1999). Statistical analysis using bootstrapping: concepts and Implementation. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research.* (pp. 81-103). Thousand Oaks, California: Sage.

Ontvangen: 14 december 2000
Geaccepteerd: 3 mei 2001