Development of metacognitive skills in young adolescents : a bumpy ride to the high road

Stel, M. van der

Chapter 6

Methodological issues

Although calls for longitudinal designs in developmental studies are frequently made, the vast part of developmental studies consists of studies with a cross-sectional design. "Although the goal of developmental psychology is to understand change in behavior across the life span, most empirical work provides only a short snapshot view of behavioral change, because most developmental research infers change by comparing *different* people of *different* ages (cross-sectional studies) rather than attempting to follow change within the same people over time (longitudinal studies studying intra-individual changes)" (Bullock & Schneider, 2009, p. 1). One of the most important reasons for this lack of longitudinal studies is a pragmatic one. Longitudinal studies not only require an investment in time and resources, but the researcher also has to deal with the risk (and the possible consequences) of participants dropping out before data collection has been finished. In this chapter some methodological issues related to the longitudinal design of this thesis are discussed.

6.1 Participants

At the start of the data collection in 2005, 32 first-year secondary-school students were included in the study. They were recruited from 85 students (for further details see Chapter 2, sections 2.2.1 and 2.2.2). Unfortunately, the repeated presentation of two tasks each year for three consecutive years and the very time-consuming, and labor-intensive analysis of thinking-aloud protocols of individual student sessions did not allow for a larger sample. All students came from the same school located in one of the largest cities in the Netherlands. This particular school is known because of its large diversity of children, thus representing a broad educational level of students, a broad range of social economic status of parents, and various ethnic backgrounds. The choice for just one school was deliberately made to avoid confounding variables, such as differences in teachers, pedagogical and/or didactic philosophy, schoolbooks, etc.

As stated above, a longitudinal design always risks an early dropout of participants, resulting in an unrepresentative sample. Loss of participants also happened in this study. After the first year, four participants dropped out due to a change of residence or school. From the remaining 28 students in the second year, three students were lost after the second year for the same reasons. This resulted in 25 participants in the third year. As long as there is no selective drop out, the loss of participants will not necessarily affect the results. Therefore, several checks were performed in order to assess any effect that might have resulted from the dropout.

In Table 6.1 descriptives for the first-year participants are shown. Intellectual-ability scores were transformed into z-scores, resulting in a mean of 0 and a standard deviation (*SD*) of 1. Participants, however, were equally distributed over the four quartiles. Unfortunately, boys and girls were not equally represented (12 boys, 20 girls), but no effect of gender was found. The average age was 12 years and 8 months. By calculating the skewness (see Table 6.1) of the independent variables (intellectual ability and metacognitive skills), it was investigated whether the scores on intellectual ability and metacognition were divided symmetrically. Results of correlational analyses and t-tests would be affected by a skewness larger than 1 (De Vocht, 2009).

Table 6.1 Descriptives participants 1st year

1st year *N*=32	Intellectual ability	Meta QLHis	Meta QLMath	Meta QNHis	Meta QNMath	LP Hist	LP Math
Mean (*SD*)	0.00 (1.00)	5.03 (3.40)	6.09 (3.30)	17.34 (12.83)	21.75 (8.60)	8.46 (3.28)	4.43 (2.53)
Skewness	.44	.36	.80	.85	.43		

Note: MetaQLHis means qualitative scores for metacognitive skills in history; MetaQLMath means qualitative scores for metacognitive skills in math. MetaQNHis means quantative scores for metacognitive skills in history; MetaQNMath means quantitative scores for metacognitive skills in math; LP means Learning performance.

Although the skewness of intellectual-ability and metacognition scores was smaller than .85, scores of both variables were further checked by transforming them into square root scores. According to Tabachnick and Fidell (2007), such a transformation would neutralize a skewed distribution. Transformed scores, however, did not alter the results of correlational analyses (see Chapter 2, Tables 2.3 and 2.4).

In order to find cases with exceptional values that would have a strong and undesirable impact on the results, Leverage values and Cook's distance were calculated (De Vocht, 2009). All cases had a Leverage value smaller than 0.5 and a Cook's distance smaller than 1, meaning that no cases had an extreme impact on the results. Therefore, no participants were to be excluded.

The same procedure was applied to the data of the second year and the third year. Again the skewness of both intellectual ability and metacognition was smaller than 1, and all cases had a Leverage value smaller than 0.5 and a Cook's distance smaller than 1. Transformation into square root scores did not lead to other conclusions.

Furthermore, the participants that dropped out after the first year and the second year have been examined more closely. After the first year, two boys (one out of the 4th quartile of intellectual-ability scores and one out of the 3rd quartile) and two girls (both out of the 4th quartile) were lost. None of these dropouts had the highest or lowest scores on intellectual ability, metacognitive skills, or learning performance. It was also investigated whether the results of the dropouts differed significantly from those of the 'stayers'. Therefore, ANOVAs were performed on the scores on intellectual ability, metacognitive skills, and learning performance, contrasting the dropouts with the rest of the participants. No significant differences were found. Because the number of dropouts was only four and their scores were not normally distributed, Mann-Whitney tests were performed too. This test is a non-parametric test without requirements of the sample distribution**.** No significant differences were found either (see Table 6.2).

Table 6.2 Results ANOVA and Mann-Whitney comparing results dropouts and 'stayers' after the first year

***N*=32**	***F***	**Sig**	**Asymp. sig**
Intellectual ability	3.51	.07	.07
MetaQLHis	.23	.63	.71
MetaQNHis	.00	.99	.75
LPHis	.02	.89	1.00
MetaQLMath	.49	.49	.55
MetaQNMath	.30	.58	.77
LPMath	.33	.57	.66

Note: MetaQLHis means qualitative scores for metacognitive skills in history; MetaQLMath means qualitative scores for metacognitive skills in math. MetaQNHis means quantative scores for metacognitive skills in history; MetaQNMath means quantitative scores for metacognitive skills in math; LP means Learning performance.

In conclusion, there was no indication of selective loss of participants after the first year. Both ANOVAs and Mann-Whitney tests did not reveal any significant differences between dropouts and 'stayers'. Moreover, the loss of four participants from the sample resulted in a perfect skewness (0.00) for intellectual-ability scores of the second year. In the second year no effect of gender was found either.

After the second year, another two boys (one out of the 1st quartile of intellectual-ability scores and one out of the 3rd quartile) and one girl (out of the 1st quartile) were lost. Again no effect of gender was found. Tests for significant differences between the dropouts and the rest of the participants were repeated on the data of the second year (see Table 6.3).

Again ANOVA nor Mann-Whitney tests revealed a significant difference between dropouts and 'stayers'. The participants that dropped out after the second year did not have any extreme scores either. These tests lead to the conclusion that participants, who dropped out of the sample after the first year and the second year, did not cause a selective loss and, therefore, did not affect the results of the longitudinal study.

Table 6.3 Results ANOVA and Mann-Whitney comparing results dropouts and 'stayers' after the second year

N=28	*F*	Sig	Asymp. sig
Intellectual ability	.96	.34	.35
MetaQLHis	1.64	.21	.15
MetaQNHis	.69	.41	.33
LPHis	2.11	.16	.12
MetaQLMath	.69	.41	.50
MetaQNMath	.82	.37	.30
LPMath	1.57	.22	.17

Note: MetaQLHis means qualitative scores for metacognitive skills in history; MetaQLMath means qualitative scores for metacognitive skills in math. MetaQNHis means quantative scores for metacognitive skills in history; MetaQNMath means quantitative scores for metacognitive skills in math; LP means Learning performance.

6.2 Intellectual ability and Metacognitive skills

6.2.1 Intellectual ability

Although intellectual ability generally is considered as a rather stable person-related feature at the age of 13, intellectual ability was repeatedly measured each year. In order to be able to answer the research question whether intellectual ability and metacognitive skills develop in line with the monotonic development hypothesis, a measure for *growth* of intellectual ability was needed. Raw scores on the various subtests were used, because norm scores would have ruled out age effects. Although there was a period of one year between two consecutive assessments, it could be argued that a repeated measure of intellectual ability involves a test-retest effect. However, no substantial intra-individual changes in test scores were observed: Over the years 85% of the participants remained in the same quartile or at the boundary between two quartiles. Therefore, it is not likely that the repeated assessments of intellectual ability yielded a test-retest effect that would distort the results of the correlational analyses (see Chapter 2, Tables 2.3 and 2.4, Chapter 3, Table 3.3, Chapter 4, Table 4.2).

To establish whether intellectual-ability scores over the three consecutive years represented the same construct, correlations of intellectual-ability scores over the three years were calculated (see Table 6.4) and a principal component analysis (PCA) was performed on subscale level (see Table 6.5; see Appendix C for correlations on subscale level).

Table 6.4 Correlations of Intellectual ability over the years

N=25	IA13	IA14
IA13		
IA14	.75**	
IA15	.71**	.82**

Note: IA13 means Intellectual ability at 13 yr. ** $p < 0.01$

Table 6.5 PCA on subtests of Intellectual-ability tests over the years

N=25	1st component	2nd component
Eigenvalue	5.59	2.01
Variance proportion	.47	.17
Number series13yrs	.70	-.23
Number series14yrs	.83	-.07
Number series15yrs	.75	.06
Verbal analogies13yrs	.69	-.44
Verbal analogies14yrs	.52	-.69
Verbal analogies15yrs	.57	-.51
Unfolding figures13yrs	.82	-.04
Unfolding figures14yrs	.83	.29
Unfolding figures15yrs	.78	.09
Memory13yrs	.50	.47
Memory14yrs	.46	.64
Memory15 yrs	.59	.55

Correlations in Table 6.4 are rather high varying from .71 to .82. Furthermore, the principal component analysis (PCA) on intellectual-ability subtest scores of all three years resulted in a two-factor solution with eigenvalues of 5.59 of the first component and 2.01 of the second component, with 47% and 17% of variance accounted for respectively (see Table 6.5). The first component could be interpreted as representing general intellectual ability and the second component could be interpreted as representing memory ability contrasted with verbal-analogical reasoning. These results allow for considering intellectual ability as a stable person-related characteristic over the years. The rather high correlations of intellectual-ability scores over the years together with the two-factor solution allow for the conclusion that intellectual ability can be considered as the same construct over the years.

6.2.2 Metacognitive skills

The transcribed thinking-aloud protocols for both tasks were analyzed on spontaneous use of metacognitive skills according to the procedure of Veenman (1993; Veenman & Beishuizen, 2004; Veenman, Kerseboom, & Imthorn, 2000). For a description of the scoring method, see Chapter 2, section 2.2.4 and Appendices A and B). Although method and rater were the same for the three consecutive years, it can be argued that a difference in rating metacognitive skills might have occurred during the study. This would mean that a certain activity was rated differently from one year to another, resulting in metacognitive skillfulness as an unstable construct. Therefore, after the third-year protocols were rated, six protocols of each task and year (36 protocols) were rated again on the same day. In order to check the consistency in rating throughout the years, a paired-samples t-test on prior scores and most recent scores was performed. No significant difference in rating between the two moments of rating was found ($t(35) = .33$, $p = .74$), meaning that quality of metacognitive skills was not rated differentially. The correlation between ratings of the two moments was very high (.99), meaning that participants were rated consistently.

6.2.3 Relation between intellectual ability and metacognitive skills

So far, intellectual ability and metacognitive skills were discussed separately in this chapter. From a developmental perspective, it is relevant to establish whether the correlations of these two independent variables with learning performance are stable over the years. In the graphs below, the correlation of each independent variable with learning performance is shown for math and history (see Figures 6.1 and 6.2).

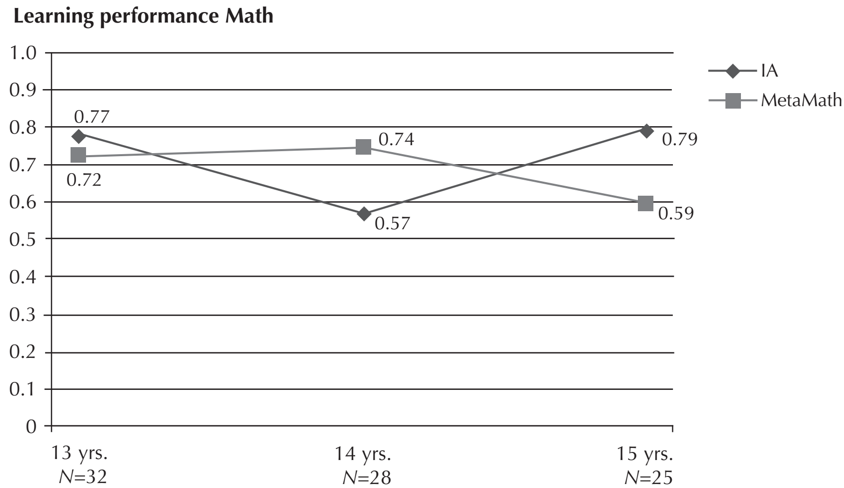

Figure 6.1 Correlations between Intellectual ability (IA), quality of metacognitive skills in math, and Learning performance Math

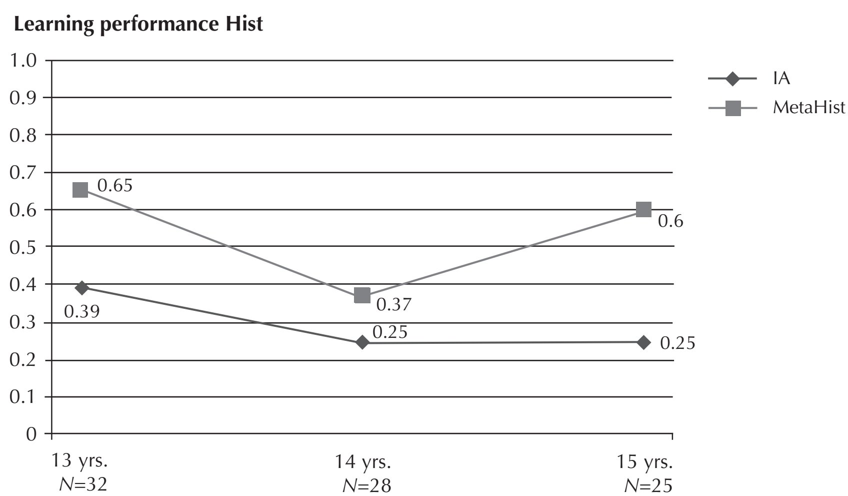

Figure 6.2 Correlations between Intellectual ability (IA), quality of metacognitive skills in history, and Learning performance History

Despite correlations seem to differ at face value over the years, Fisher-*z* ratios (see Table 6.6) show that there is no significant difference between the correlations over the years. All Fisher-z ratios are smaller than 1.96 (Guilford, 1965).

Table 6.6 Fisher-*z* ratios

Math	13yrs(32)	14yrs(28)	F*z*	14yrs(28)	15yrs(25)	F*z*
IA-LP	.77	.57	1.37	.57	.79	1.45
MS-LP	.72	.74	.16	.74	.59	.93
IA-MS	.62	.51	.60	.51	.66	.79
History	**13yrs(32)**	**14yrs(28)**	**F*z***	**14yrs(28)**	**15yrs(25)**	**F*z***
IA-LP	.39	.25	.57	.25	.25	.00
MS-LP	.65	.37	1.42	.37	.60	1.04
IA-MS	.27	.32	.82	.32	.41	.36

Note: IA means intellectual ability; LP means learning performance; MS means Metacognition (qualitative scores). (32) means *N*=32; (28) means *N*=28; (25) means *N*=25; F*z* means Fisher-*z* ratio.

In the graphs below (see Figures 6.3 and 6.4), the unique contribution in percentage of both intellectual ability and metacognitive skills for both tasks separately are shown. The shared variance accounted for by both variables share is shown as well.

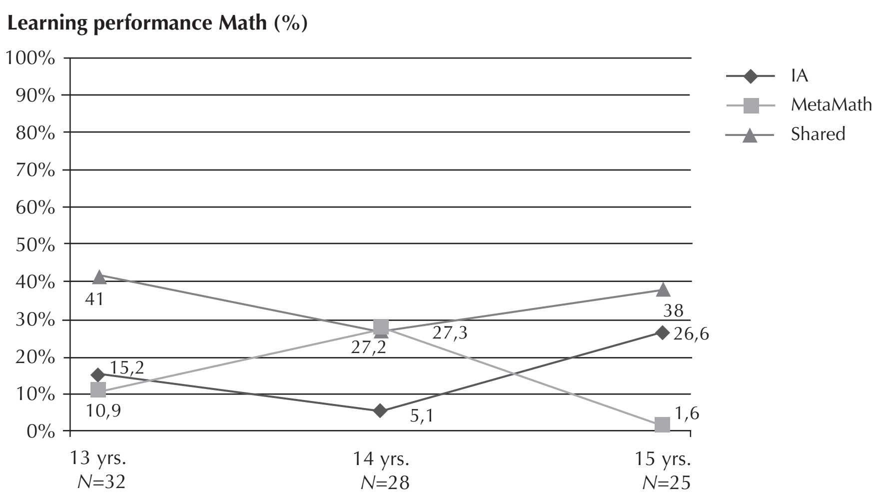

Figure 6.3 Distribution of unique and shared variance accounted for in Learning performance Math

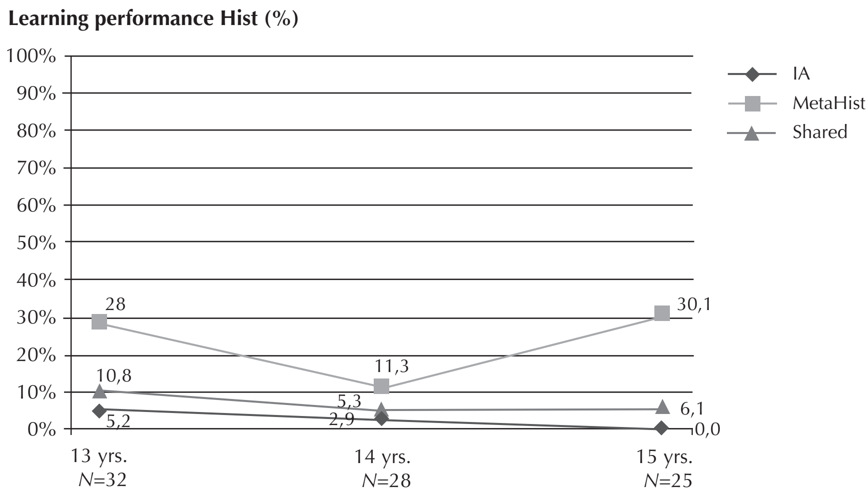

Figure 6.4 Distribution of unique and shared variance accounted for in Learning performance History

To calculate the unique and shared contribution of intellectual ability and metacognitive skills to learning performance, depicted in Figures 6.3 and 6.4, semi-partial correlations are needed. If semi-partial correlations would differ significantly, the stability of these correlations would be at risk. This would cause instability in the unique variance accounted for of either one of the predictors of learning performance. Therefore, Fisher-*z* ratios were calculated for the semi-partial correlations as well (see Table 6.7).

Table 6.7 Fisher-*z* ratios for Semi-partial correlations

Math	**13yrs(32)**	**14yrs(28)**	**F*z***	**14yrs(28)**	**15yrs(25)**	**F*z***
SemIA-LP	.39	.22	.66	.22	.52	1.15
SemMS-LP	.33	.52	.82	.52	.13	1.46
History	**13yrs(32)**	**14yrs(28)**	**F*z***	**14yrs(28)**	**15yrs(25)**	**F*z***
SemIA-LP	.23	.17	.22	.17	.00	.56
SemMS-LP	.53	.33	.87	.33	.55	.90

Note: SemIA means semi-partial correlation intellectual ability; LP means learning performance; SemMS means semi-partial correlation Metacognition (qualitative scores). (32) means *N*=32; (28) means *N*=28; (25) means *N*=25; Fz means Fisher-*z* ratio.

Again all Fisher-*z* ratios were smaller than 1.96. This means, for instance, that it cannot be concluded that the semi-partial correlation of IA and LP at the age of 15 (.52) is higher than the one at the age of 14 (.22). It can be concluded that the correlations and the semi-partial correlations were stable over the years.

6.3 Learning tasks

In this thesis, intellectual ability and metacognitive skills were investigated as predictors of learning performance. In order to investigate the unique contribution of metacognitive skills to learning performance from a developmental perspective, *new* learning tasks and post-tests were required each year. In order to exclude confounding variables as much as possible, only the content of the learning tasks was adapted each year to age and grade, while the format of the learning tasks and post-tests was kept the same for both math and history.

6.3.1 Math

For math the content of assignments in the learning-by-doing phase were adapted to age and grade each year. However, the tasks were made as comparable as possible as far as it concerned the format. Each year, the tasks were ecologically valid, because they were composed of adaptations of math problems from a frequently used schoolbook for math in the Netherlands. The tasks were piloted in the age group of participants.

Although, the content of assignments had to be new each year, items with a comparable content, that is, an ascending level of difficulty within the same area of math, were included in the tasks over the years. For example, in every year a geometry assignment was included. In the first year participants had to calculate the circumference of a meadow, in the second year it concerned the surface area of a triangle, and in the last year they had to apply Pythagoras' theorem in order to calculate the horizontal side of a triangle.

6.3.2 History

To make sure that the texts were suitable for text studying and measuring learning performance afterwards, the content of the texts had to be new to the participants. Participants likely had little or no content knowledge about the topic of the text, because topics were taken from the curriculum that was one year ahead. All learning tasks were piloted as well. No familiarity with the topics was observed.

In order to minimize a possible confounding effect of differences in learning texts, the *format* of the texts was made as comparable as possible. Van Hout-Wolters (1986) described various text characteristics that affect learning processes and/or learning results in text studying: Type of text, structure, difficulty of language used, length, and didactic help. All these variables were taken into account when composing the history learning tasks. All texts were informative, and ecologically valid (derived from the two most frequently used schoolbooks in the Netherlands). Structure, layout, and length of the text were kept almost identical (see Table 6.8). In order to be suitable for text studying, texts need to be of a certain length. Texts that are too short will only be memorized instead of being studied. In each text, the same didactic help was embedded, that is, three activating questions and/or assignments were included in order to elicit (more) metacognitive activities.

Table 6.8 Text characteristics History tasks

Number of	1st year	2nd year	3rd year
Pages	4,5	5	5
Paragraphs	18	19	20
Words	1497	1518	1522
Words per sentence	12	12	12
Concepts	76	95	122

According to Van Hout-Wolters (1986), Veenman and Beishuizen (2004), and Weaver and Bryant (1995), the average number of words per sentence is an important indicator for the difficulty and readability of texts. Each year, the average number of words per sentence was twelve. The other text variables depicted in Table 6.8 remain almost the same as well, except for the number of concepts. It is assumed that the texts are equivalent as far as the format was concerned. The content, on the other hand, can be considered as more abstract and more 'condensed', thus more difficult, because of the growing number of concepts introduced in the texts.

6.4 Post-tests

As explained above, new post-tests were needed each year to measure the learning performance. The content of the post-tests had to be adapted to the learning tasks and had to be suitable for the age group. Like the learning tasks, the post-tests were piloted before they were administered to the participants. Again the format was kept the same. For math, the items in the learning phase were parallel to the items in the post-test, that is, the surface structure of the post-test items differed from the ones of the learning-task, but the deep structure was the same. For history, each post-test consisted of five multiple-choice questions and six essay questions.

6.4.1 Math

Despite the efforts to make the post-tests as comparable as possible in relative level of difficulty, participants could have perceived a difference in level of difficulty, other than just a relative difference. Moreover, the math pilot revealed that for older students more assignments were needed in a 50 minute-session relative to younger students. This resulted in a difference in the number of items and in obtainable scores over the years for math. Therefore, all post-test results were checked on differences in level of difficulty. First, the mean proportion of right answers (*p*-value) was calculated for each task (see Table 6.9).

Table 6.9 Mean P-values post-tests Math (*SD*)

***N*=25**	**1st year**	**2nd year**	**3rd year**
Math	.44 (.23)	.69 (.15)	.66 (.15)

ANOVA was performed on the p-values of all math items. A slightly significant difference was found ($F(2,22) = 3.59$, $p = .048$). Therefore, a post-hoc test (Bonferroni) was performed on the math data. This post-hoc test did not reveal a significant difference between the learning outcomes for math over the three years.
Correlations between learning performance measures over the years vary from .63 to .66 (see Appendix C).

6.4.2 History

In history, correlations between learning performance measures over the years vary from .48 to .74 (see Appendix C). ANOVA was repeated for the *p*-values of the history post-tests (for mean p-values, see Table 6.10).

Table 6.10 Mean P-values post-tests History (*SD*)

N=25	1st year	2nd year	3rd year
History	.51 (.26)	.54 (.29)	.56 (.30)

ANOVA on the p-values of all history items did not reveal a significant difference ($F(2,37) = .11$, $p = .90$). In conclusion, no significant difference in the relative difficulty level of the post-tests for math and history was found.

6.5 Testing generality vs. domain specificity of metacognitive skills

To establish whether metacognitive skills are general or domain specific by nature, principal component analysis (PCA) was performed, each year separately. For these analyses unrotated PCAs were used. An unrotated solution is a so-called direct solution, that is, originating directly from correlational matrices, while derivated solutions result from rotation and are obtained from direct solutions. In this thesis, no use was made of any rotation, because a rotation would rotate out the general component. Justification for this decision was found in the literature on factor analysis. If you are looking for separate, contrasting factors and want to avoid a general factor, you need to rotate (e.g., for the purpose of test validation). A rotation, however, should *not* be used if the theoretical expectation suggests that a general factor may occur (Gregory, 1996). Unrotated solutions are frequently applied in exploratory factor analysis, summarizing interrelationships between a number of variables. In this thesis, PCAs were performed to find an answer on one of the research questions, that is, to establish whether metacognitive skills are general or domain specific by nature. Using PCAs as a tool to address this issue was in line with prior studies (Veenman & Beishuizen, 2004; Veenman & Verheij, 2003; Veenman, Wilhelm & Beishuizen, 2004). It certainly would have been interesting to perform confirmatory factor analysis as well (see Veenman, Elshout, & Meijer, 1997). Then the purpose is to confirm that variables fit a certain pattern predicted by a theory. Lisrel and SEM are examples of such model-testing programs. Unfortunately, the sample was too small to do so. Hayduk (1987) stated that goodness of fit can be estimated for small samples, but only in experimental designs (cf. Veenman et al., 1997). The studies in this thesis, however, do not have an experimental design.

In this chapter, several checks were performed in order to rule out confounding effects due to the longitudinal design. In summary, loss of participants did not cause a selective loss of participants, repeated measures of metacognitive skills did not lead a significant difference in rating between the moments of rating, intellectual ability can be considered as the same construct over the years, and no significant difference in the relative difficulty level of the post-tests for math and history was found.