Self-Regulated Learning
as a Cross-Curricular Competency

The Construction of Instruments in 22 Countries
for the PISA Main Study 2000

Synthesis Report

Jules L. Peschar
René Veenstra
Ivo W. Molenaar

with
Anne Boomsma
Mark Huisman
Marieke van der Wal

University of Groningen
Department of Sociology
Department of Statistics, Measurement Theory and Information Technology

Presented at the Network A Meeting in Echternach, Luxembourg

October 27-29, 1999

American Institutes for Research (AIR)
3333 K St., NW
Washington DC 20007, USA

Table of contents

Synthesis Report

	Executive Summary	5
1	Introduction	7
2	Goals of the Analysis	7
3	Organisation, Role of the International Experts and National Feedback	8
4	Potential and Limitations of the Data	9
5	Strategy of the Analysis	12
	5.1 Definition and Reporting of Scales	12
	5.2 Strategy	13
	5.3 Feedback from National Sources	15
6	The Scales	16
	6.1 General Remarks	16
	6.2 Learning Strategies	18
	6.3 Motivational Preferences	20
	6.4 Goal Orientation	23
	6.5 Self Related Cognitions	25
	6.6 Action Control: Effort and Persistence	28
	6.7 Learning Preferences	30
	6.8 Implicit Theories of Learning	31
7	Conclusions of the Analysis	33
8	Quality of the Scales for PISA 2000	36
	8.1 Reliability of the Scales	36
	8.2 Construct Validity	37
	8.3 Related Issues	40
9	Presentation of the Scales	41
	References	44

Technical Report

1 Introduction
2 Descriptive analyses
3 Reliability analyses
4 IRT analyses MSP
5 Analyses for subgroups: descriptives
6 Analyses for subgroups; reliabilities
7 Correlations of scales with Achievement Tests
8 Effects of different estimation methods

Executive Summary

Self-regulated Learning

Data on cross-curricular competencies (CCC) and self-regulated learning were collected in 22 countries in spring 1999, according to the plan described by Baumert, Fend, O'Neil, and Peschar (1998). Different versions of theoretically relevant scales were selected and used to collect the data. On the basis of statistical analysis and expert judgement, the most appropriate instruments have been selected for use in PISA in 2000.

This report discusses the analysis of these data, according to the procedures described in the Plan of Analysis of Self-Regulated Learning as a Cross-Curricular Competency in PISA's Field Study (Peschar and Molenaar, 1999) as it was accepted by the OECD and its Technical Advisory Group for PISA. In selecting items, preference has been given to items that have good or excellent psychometric properties. In addition, items that have good conceptual coverage and have a clear relevance to educational policy makers are favored. In consequence, the final instrument for measuring self-regulating learning has fewer than the original 112 items (and 23 scales) included in the 1999 field test. The recommended instrument has 51 items and 14 scales to distinguish three self-regulated learning dimensions: Learning Strategies, Motivation and Self-Concept. The recommended scales have good reliability and validity. The items selected can be successfully administered to students within no more than 10 minutes.

Strategy of Analysis

The research team applied classic reliability tests, confirmatory factor analysis (CFA), and item response theory (IRT). The analysis focused on confirmatory procedures and reduced the number of scales or items in an optimal way. Scales and/or dimensions were deleted or re-arranged primarily on the basis of two criteria: psychometrical performance and content arguments.

An international expert group (J. Baumert, H. Fend, H. Marsh, H. O'Neil, L. Van de Poele and J.L. Peschar,) critically reviewed the statistical analysis of the field-trial data and made recommendations on the selection of items and construction of the final instrument. The expert group also advised the research team on the selection of items and scales on the basis of psychometric, theoretical, and policy relevance. The experts gave special attention to the issue of differential item functioning and validity of the scales. The experts reviewed various drafts of this report and endorsed this final report.

Criteria for Selecting Scales

The scales were selected according to their theoretical, statistical, political and educational relevance. Several statistical models were used to assess the scales' psychometric properties: descriptive analyses, reliability estimates, correlations, CFA, structural equation models, and nonparametric IRT scaling. The scales selected meet the standards of these modeling techniques.

Furthermore, the scales were evaluated on theoretical and content issues, such as:

- A dimension should contain more than one scale, to prevent a lack of balance
- Overlap between the scales should be avoided
- The measured construct should be teachable
- The construct measured should be valuable in regard to the curriculum
- The construct measured should be amenable to influence through policy interventions

Selection of Scales

The report concludes that PISA should include three main dimensions of self-regulating learning, namely: (a) learning strategies; (b) motivation; and (c) self-concept. These dimensions characterized the final set of scales more appropriately than the original set of seven dimensions. Scales selected in these domains are content relevant and meet or exceed the psychometric standards established for the analysis.

The following scales are suggested as the Self-Regulated Learning Instrument:

- **Learning strategies: scales on memorising, elaboration, and control strategies;**
- **Motivation: scales on instrumental motivation; interest motivation subject related, action control; effort and persistence in learning; cooperative and competitive learning.**
- **Self-concept: control expectation, self-efficacy, verbal self-concept, self-concept in math, and overall academic self-concept.**

General Suggestions

The expert group and the research team strongly recommend that the CCC questionnaire be administered **before** the achievement tests. Such an ordering will prevent a situation in which students' perceptions about how well (or poorly) they performed on the achievement tests affect responses to the items about self-regulated learning. We need to avoid a situation in which attitudes about the achievement tests might provide a possible explanation for responses to items about self-regulated learning.

Both the research team and the expert group advise against the use of single items to represent entire scales or domains. The reliabilities and psychometric properties of such items are unknown and likely to be poor.

It is strongly recommended that a single **core** questionnaire be used (i.e., that rotation of items **not** be used). For modeling purposes, all scales and items for all students should be available so that correlations among all scales and items can be determined in the final analysis of the data of the main study.

Finally, this analysis suggests that the translation process needs to be monitored intensively to achieve highly equivalent instruments.

1. Introduction

In the early 1990s several members of Network A suggested the desirability of broadening the scope of the educational indicators in the INES programme. It was observed that the typical indicators of educational achievement could be supplemented with information about competencies that are not taught within a particular subject at school.

These so-called cross-curricular competencies (CCCs) were the focus of a developmental trajectory between 1993 and 1997. Measurement in four domains -- civic knowledge, problem solving, self-cognition, and communication skills -- was piloted in nine countries. The results were promising. The studies demonstrated that competencies in civic knowledge and self-cognition can be measured with existing instruments of good quality (OECD, 1997). For problem solving and communication, no satisfactory results were obtained.

On the basis of these results, Network A -- and later the INES steering group and other OECD bodies -- decided to offer two of the CCC domains as optional supplements to PISA. Items to assess self-cognition will be an option for PISA's administration in 2000. To date, 20 countries have indicated their intention to assess self-cognition.

In preparation for the inclusion of items to assess self-cognition, a panel of experts refined the definition of CCCs, which includes the following concepts (see Baumert et al 1998):

- Strategies of self-regulated learning, which regulate how deeply and how systematically information is processed;
- Motivational preferences and goal orientations, which regulate the investment of time and mental energy for learning purposes and influence the choice of learning strategies;
- Self-related cognition mechanisms, which regulate the standards, aims, and processes of action;
- Action control strategies, particularly effort and persistence, which help to concentrate on the performance of actions and help to overcome learning difficulties; and
- Preferences for different types of learning situations, learning styles, and social skills required for co-operative learning.

2. Goals of the Analysis

PISA's field test has generated data on self-regulated learning. The field test included a series of items, in three nonoverlapping "packets," that asked students to indicate their opinions and attitudes about learning. The present report addresses issues with respect to the analysis and further use of the items during PISA's administration in 2000. One important goal of the analysis has been to reduce the number of items to about one third -- to comply with the available testing time -- and to establish the measurement properties for the instruments to be applied for the next few years.

This report contains the following information on outcomes of the field study:

- The definition of the scales that the analysis plan aims to construct and a proposal for how the resultant scores will be reported;
- The methodology used to create the scales;
- A proposal for the selection of the final items and construction of the instrument.
- The methods and criteria for assessing construct validity; and
- The methods and criteria for assessing the reliability of the scales.

3. Organization, Role of International Experts and National Feedback

A research team at the University of Groningen's Departments of Sociology and Statistics, Measurement Theory and Information Technology analyzed the data. Members of the team include:

René Veenstra	Post Doctoral Fellow, Department of Sociology
Ivo W. Molenaar	Professor of Statistics and Measurement Theory
Anne Boomsma	Associate Professor of Statistics and Measurement Theory
Mark Huisman	Research Associate, Statistics Netherlands, Voorburg
Marieke van der Wal	Research Associate, Department of Sociology
Jules L. Peschar	Professor of Sociology of Education, Project Director

A group of **international experts** convened in Groningen on September 27-28, 1999, and met with the research team to discuss the results of the statistical analysis and to offer advice and suggestions to the research team. The expert group included:

Harold F. O'Neil	Professor, University of Southern California, Los Angeles, USA
Herbert S. Marsh	Professor, University of Western Sydney, Macarthur, Australia
Luc Van de Poele	Senior Research Associate, University of Ghent, Belgium
Juergen Baumert	Professor, Max Planck Institute for Education Research, Berlin, Germany, represented by Dr. Cordula Artelt
Helmut Fend	Professor, University of Zürich, Switzerland (apologies)
Jules Peschar	Professor, University of Groningen, the Netherlands, Project Director
Richard Tobin	Principal Research Scientist, American Institutes for Research, Washington, D.C., USA

The experts reviewed draft versions of the instruments critically, made suggestions for improvement, and made recommendations on the selection of items and construction of the final instrument. Subsequent to the meeting of the expert group, the research team distributed a summary report and solicited comments from each member of the expert group. This report reflects and incorporates the comments received from the experts.

In addition, **national feedback** was sought, parallel to the consultations with the group of experts. National project managers (NPMs) and those involved in the national CCC preparations were consulted in two ways. First, they were invited to report on the administration of the field trials. This is a relevant source of information on the functioning of items as well as on translation issues; the results are reported in Section 5.3. Second, as a result of the initial analysis, NPMs will

be asked to provide feedback on translation issues and on the proposed instruments in order to evaluate and select reliable scales or acceptable differences among subpopulations in their countries. These consultations are intended to promote national support for the instruments on self-regulated learning.

4. Potential and Limitations of the Data

The data on which this report draws were collected in accordance with the Framework of the PISA Pilot Study. Due to the limited testing time available in the pilot, not all 112 items initially selected (Baumert *et al*, 1998) could be presented to all students involved. Therefore, another strategy was followed. Three nonoverlapping packets, each with approximately one third of the total pool of items, were composed and presented to three different samples. The packets used in the field trial did not have similar conceptual coverage, as can be seen in Table 1.

Table 1: Overview of Packets

Packet (# of items)	**Dimensions of Self-Regulated Learning**		
Sample	**Packet 1** (Q1, Q2) (41 items**)**	**Packet 2** (Q3, Q4) (42 items**)**	**Packet 3** (Q5, Q6) (29 items**)**
Sub Sample 1	• A: Learning Strategies(17) • B: Motivational Preferences (24)		
Sub Sample 2		• C: Goal Orientation (9) • D: Self-Related Cognitions(33)	
Sub Sample 3			• E: Action Control: Effort and Persistence (10) • F: Leaning Preferences (8) • G: Implicit Theories of Learning (11)

Note: Q1. Q6 refers to parts of the questionnaire used in the field trial
A..G refers to the dimensions distinguished. Within dimensions an additional digit is given to each scale.

As a consequence of this approach, the search for optimal instruments was conducted *within* each of the packets. There was *no overlap* between samples and packets, so it was not possible to conduct analyses on the combined samples and merge items from different packets into one analysis. This is a pity from an analytical point of view, but in the analysis stage we cannot repair decisions that were taken in the data-collection stage.[1] We devote attention to this issue and the consequences it has for our conclusions below (Section 7)

The data were collected in 22 countries, namely USA, Canada (French and English versions), Mexico, Brazil, Australia, New Zealand, Korea, Russian Federation, Austria, Belgium, Czech Republic, Denmark, Finland, Germany, Hungary, Ireland, Italy, Luxembourg, Netherlands, Norway, Sweden, Switzerland (French, Italian and German versions). Since the purpose of the

[1] This is especially relevant for Competitive Motivation (scale B3) and Learning Preferences (scale F2).

present report is to construct instruments, rather then explaining country differences, the countries are relabeled. For Buttercup and Iris the data could not be included in the latest analyses, though they have been collected.

Table 2: Countries involved

Country	Available[2]	
Anemone	1	
Aster	1	
Buttercup	2	
Carnation	1	
Cornflower	1	
Crocus	2	
Cuckooflower	2	
Daffodil	1	
Dahlia	2	
Daisy	2	
Dandelion	2	
Fresia	1	
Gladiola	1	
Iris	2	
Jasmine	1	
Lily	1	
Poppy	1	
Rose	2	
Snowdrop	1	
Sunflower	2	
Orchid	2	
Magnolia	2	
Tulip	2	
Honeysuckle	2	
Violet	2	
Total		

Samples

No children from higher social strata (fathers' education) were included in the samples from Fresia and Anemone; no children from lower social strata (fathers' education) were included in Lily's sample. In Snowdrop no information on the socio-economic index ISEI was available. Such samples may affect differential item functioning.

[2] First wave data were made available by September 1st, second wave data by October 15th. See also section 5.2.

Background characteristics

About 75 percent of the students with CCC data completed the reading test. About one third completed the mathematics or science test. Unfortunately, students who completed the third part of the CCC questionnaire did not complete the math or science test. About two-thirds of the students provided information on their socioeconomic background.

5. General Strategy of Analysis

5.1 Definition and Reporting of the Scales

The three packets used in the field trial contained 112 items (in 23 scales) and addressed the seven dimensions of self-regulated learning identified in Table 1. Thus each scale was intended to measure a single, unidimensional construct with adequate reliability and generalizability over different countries. The overall intent of the analyses was to evaluate these psychometric properties on the basis of the Field Study data.

The structure of the scales was, therefore, predefined. The research team tried to reconstruct the intended structure from the data. As mentioned elsewhere, however, it may be the case that some dimensions cannot be reconstructed completely or satisfactorily because of the nonoverlapping nature of the three packets. Under these circumstances**,** a decision had to be made whether to retain the scales or to reject their use in PISA.

Such choices have been based on the psychometric results, but theoretical arguments have played a major role as well. Though in Baumert *et al* (1998) dimensions and instruments are not given a specific weight, some dimensions or instruments may be theoretically more relevant than others. Or, some instruments may be more relevant to education. The trade-off in such a situation was explicitly discussed with the panel of experts

When choices were required, preference was given to:

- **few instruments that have good or excellent psychometric properties;**
- **broad (theoretical) coverage of the domain of self-regulated learning; and**
- **instruments that have a clear relevance to educational practice and policy makers.**

Scores on scales were calculated on the basis of simple counting of values of answering categories. This facilitates the analysts' work as well as policy makers' interpretation of the results. In separate analyses we compared simple scoring methods with more advanced methods to assess the consequences of different methods of calculating the composite scores and to trace potential risks of this practice (see Section 8.4). It appears that our assumptions were correct.

Reliability and ease of interpretation governed our consideration of these issues. Since the scales are composed of a different number of items, the scales can easily be transformed to a range of, say, 0 to 100. Having noted this process, our intent was to adhere as closely as possible to PISA's overall scoring practices and provide scores that are analogous to the scores in reading achievement. A principal aim of the analysis was to keep the instrument uniform across countries and subgroups.

Variables available

For the analyses conducted, the following information was available:

- Scores on packets of items (for one third of students)
- Background variables (for all)
  - Gender
  - SES/Parents' education
- Criterion variables (for all)
  - Achievement scores on reading, mathematics, and science

The ISCO classification was used to code the job of the parents as an indicator for SES. This information was then converted to the International Socio-Economic Index (ISEI). This information is available for the mother and the father. We used fathers' SES. In the field trial, achievement measures were collected in nine different booklets. The Australian Council for Educational Research (ACER) standardised the achievement scores to ensure their comparability through booklets and by countries.

Subject-related attitudes on self-perception in mathematics, reading, or science were also collected during the field trial. These attitudes might have been used for extra validation purposes. Similar attitudes were also used, however, in the present analysis and thus there was no special need to do this.

5.2 Strategy

Setting the scene

ACER provided approximately 27800 responses from 15-year old students from 22 countries. The data were made available in different waves, due to the extensive cleaning procedures. Thus our strategy has been to analyze the data of the first wave of 11 counties -- comprising about 12000 students -- and to select the instruments on the basis of the first wave data. When the data from the second wave became available, the properties of the scales for these countries were calculated and compared. There were no countries where already selected scales had such obvious different statistical properties that these had to be rejected later.

Converging evidence

Both substantive and psychometric considerations were used to achieve satisfactory solutions. Nonetheless, there is no single psychometric model that is ideally suited to handle the range of subproblems associated with the two tasks noted immediately above. As stated in the analysis plan, we used:

- Cronbachs' alpha, means, standard deviations, and correlations (using SPSS);
- Structural equation models (factor analysis and subgroup comparison) (using LISREL);
- Confirmatory/exploratory scaling within nonparametric IRT (Using MSP); and,
- Where relevant and possible, logistic IRT modeling (using ConQuest/OPLM).

Through this strategy our main conclusions were guided by **converging evidence** from two or three different measurement models and thus were not dependent on the model choice, about which methodologists may not always have the same preference.

Attitude versus achievement measurement

In the past 30 years, an elaborate testing technology has been developed. Its statistical models strive for elegant answers to subtle questions, to be derived from complicated data structures. This technology has been developed primarily for measuring educational achievement, however. If we compare this domain with the CCC domain, we must expect to be faced with somewhat less favorable circumstances. The field tests asked for *self reports* about what students feel or how they study, in a printed questionnaire answered in a classroom under time pressure at the end of a long session measuring achievement. Under these conditions:

- it is difficult to use large numbers of items that adequately tap the concepts of interest; students might be less motivated to provide honest and valid responses;

- the choice among four response categories per item has a higher risk of being influenced by other aspects than someone's value on the latent trait being measured than in the case of right-wrong scoring of achievement items;
- psychometric models for multicategory attitude measurement, although well developed and successfully used in the past, have a somewhat higher risk of exhibiting model misfit and/or low reliability.
- respondents are more likely to provide erratic answer patterns in attitude measurement than in achievement measurement.

Asking too much?

Our ideal is to measure appropriately each of the 23 scales or subdomains with due respect to the following:

- Testing time is short, thus few and simple items are preferred;
- Items should adequately and validly cover the concept of interest;
- Each subscale should be statistically reliable (preferably alpha>0.80, otherwise relations with other variables might become weak);
- The score range should permit useful discrimination (no floor or ceiling effect);
- There should be no differential item functioning that would endanger valid score comparisons among meaningful subgroups (gender, race, SES);[3]
- These requirements should apply uniformly across countries and subgroups;

Even if the answers were artificially generated by a computer exactly according to an IRT model, rather than being collected in thousands of classrooms, the statistical variability would inevitably produce some cases where some of these demands would seemingly not be met, by pure chance. It will be evident that we must be prepared to find some subscales in some countries (or in some subgroups within a country) where one or more of our quality aspects is at most "fair" rather "good" or "excellent."

Nonparametric IRT scaling with MSPWin5 Software

Mokken's (1971, 1997) proposed nonparametric IRT model as extended to multicategory items by Molenaar (1981, 1997) is closely related to work by Cliff, Ramsay, Holland, Rosenbaum, and in particular Junker and Stout (1990). The so-called MSP model incorporates exploratory and confirmatory approaches to an item pool, subgroup comparison, model checks, reliability estimation, and several other features that are useful for the CCC analysis. Once "candidate" subscales for PISA 2000 were identified, LISREL analyses were applied in order to find converging evidence. Where relevant and possible, other parametric IRT models were also applied.

The main MSP outcomes reported here are values of Loevinger's H for each scale (Loevinger, 1948). Such values provide a kind of average item-total correlation: subjects can be meaningfully ordered by their total scale score if H is at least 0.30; H>0.50 indicates a very strong scale. Contrary to Chronbachs' alpha, H does not increase with test length.

[3] Sometimes, however, differences between subgroups are meaningful and consistent with the construct. Girls have higher verbal self-concepts, whereas boys have higher math self-concepts.

When adequate measurement for each student is the primary goal, a uniform structure of the inter-item relations is desirable but not required. We have chosen monotonic homogeneity (increasing item-response functions) rather than double monotonicity (also nonintersecting item response functions) and invariant item-step ordering across groups. Checks for the latter two were examined but played a minor role in our final findings.

5.3 Feedback from National Sources

We received comments on CCC/self-regulated learning from various countries. The overall comments were positive about the CCC questionnaire. The NPMs/NPCs had few comments about the content of the questionnaire. They are the opinion that the most important concepts are covered by the questionnaire. No comments were made either on the organization of the fieldwork or the marking and data entry.

The NPMs/MPCs from the countries raised two important issues:

- The time required to complete the entire PISA questionnaires (including the subject tests) was too long. Students had to complete the student questionnaire, which contained the CCC items, at the end of the testing session. At that time most students were tired, less concentrated, and less motivated. [4]It is very likely this situation influenced the quality of the data negatively. The implication for our analyses is, however, that we will find an **underestimate of the reliabilities** of the instruments.
- The translation process has been a hasty activity in most countries. Due to this hurry, the NPMs/NPCs suspect that not all items were translated appropriately, and they argue for a check before including the items in next year's PISA questionnaire. This concern may be of relevance to the entire questionnaire, not just those items that address self-regulated learning.

Comments on specific items by countries

Several countries noted problems with the translation or content of specific items.[5] In general we interpret this national feedback as very positive. There were no serious problems with regard to the field work in the schools and apparently students were able to complete the questionnaires.

Nonetheless, due attention must be given to translation and the overload of work for students. These issues are relevant to the entire PISA questionnaire, not just those items related to self-regulated learning.

[4] The Norwegian team noted that respondents tended to respond regardless of the questions, to some extent. The Czech team noted this issue in the same manner.

[5] In the Czech Republic, as an illustration, students may have been somewhat misled by the questions related to learning because "reading" is not regarded as "learning" in this country. In Denmark there was a problem with the translation of the word "study" ("I like to study …").
In Hungary there was some difficulty with the translation of some CCC items because of doubt on where the emphasis should be put in a sentence.

6 The Scales

6.1 General Remarks

We begin our report on the findings with some general observations.

Most students completed all CCC items. Missing data appear not to be a big problem. However, the lay-out of the items is not very attractive. It is advisable to separate the CCC questionnaire into different parts to make it more appealing.

In addition, it has been noted that is probable that the correlation between items may be affected by the fact that the first part of several items is often the same (e.g., many items begin with: "When I study,…").

The CCC instrument used in the pilot study included several short scales. This may be positive from the perspective that much is covered, but the consequence is that it is thus difficult to delete items in the interest of time considerations.

Some items have as answer categories:

Type A: "Disagree – Disagree somewhat – Agree somewhat - Agree."

Others items have standard categories such as

Type B: "Almost never – Sometimes – Often – Almost always."

These categories should be clear in the questionnaires and, the categories should correspond to those used in the PISA student context questionnaires.
This will have to be coordinated in the PISA framework.

In the following sections we discuss the scales for the seven dimensions in detail. For the sake of brevity we do not repeat the theoretical justification for every item and for all scales. Baumert et al (1998) provide a complete description, so we refer to that publication.

Only relevant values of the applied psychometric criteria are discussed here. Not all values for all items are given in this report. A separate Technical Report, available upon request, provides all such values. Table 3 summarizes the methods and critical values applied in the analysis.

Table 3: Methods and Critical Values

Method	To be Improved	Good	Excellent
Reliability Chronbachs' Alpha	< .70	.70 - .80	> .80
IRT: Loevinger's H	< .40	.40 - .50	> .50

Cuttoff criteria for the evaluation of the fit of LISREL-models are not easily defined; ideally, a combination of fit indices is considered. Rough cuttoff criteria for a decision between bad and good fit are as follows: for the chi-square goodness-of-fit index a p-value smaller than 0.05 is considered to be bad, RMSEA > 0.06 (not good), SRMR > 0.08 (not good), and NNFI < 0.95 (not good). These cuttoff values are based on research of Hu and Bentler (1999).

When reliability estimates (i.e., values for Cronbach's alpha) were too low, an exploratory analysis was undertaken by deleting weak items. If this did not improve the scale, then the process of analysis was stopped. The same can be said for the IRT analyses: these were always applied. If a scale did not meet the threshold value, exploratory analyses were used to select the

weakest items for deletion. When this did not solve the problem, the analysis was stopped. In such cases, of Tentativecourse, it made no sense to check the factor structure and one-dimensionality by LISREL For a few countries a number of estimation methods for LISREL models were compared: maximum likelihood estimates, robust corrections for standard errors and the chi-square goodness-of-fit
statistic, and weighted least squares estimates. Details of these comparisons are given in the Technical Report. Our general conclusion is that the choice of the estimation methods does not affect the scale decisions.

It is important to note that single-item scales are not considered relevant because none of the statistical analyses used could produce an indication of the quality of the item. Furthermore, two-item instruments were neglected for the same reason.

It may be relevant to emphasize that the selection of the scales --as reported in the next sections—is based in the Wave 1 data for 11 countries. After the Wave 2 data became available, a separate check has been done on all 22 countries, to ensure the quality of all selected scales. [6]

[6] The details are given in the Technical Report, Section 6.

6.2 Learning Strategies

In the theoretical and developmental stage three relevant dimensions for Learning Strategies were distinguished. These items were assessed through the use of 17 items in the CCC questionnaire. If the items appropriately measure the underlying concept(s), the statistical analysis should demonstrate acceptable levels of reliability and scalability and thus justify the items' use.

A1: Memorising
A1Q1A01 - When I study, I try to memorise everything that might be covered.
A1Q1H08 - When I study, I memorise as much as possible.
A1Q1N14 - When I study, I memorise all new material so that I can recite it.
A1Q1T20 - When I study, I practice by saying the material to myself over and over.

With regard to the Memorising Scale (A1), this turned out to be only partly possible. The scale's reliability (Cronbach's alpha) is between 0.60 and 0.83 in the countries. In four countries (Fresia, Gladiola, Jasmine, and Lily) the alpha value is below 0.70. These values diminish the scale's appeal. Moreover, the formulation of the first and second items is similar, and this inflates reliabilities.

The IRT criteria, Loevinger's H coefficient, is between 0.34 and 0.58. This indicates again that the scale should be improved. Most problems center on the last item, although the content is clear and fine. In addition, the mean value of the scale varies considerably among countries.

Conclusion: not a strong scale. Only when content arguments are relevant could it be maintained. Otherwise the scale should not be used.

For the second dimension of Learning Strategies, two alternative four-item scales were suggested: Elaboration (A2-1) and Transformation (A2-2)

A2-1 Elaboration
A2Q1D04 - When I study, I try to relate new material to things I have learned in other subjects.
A2Q1K11 - When I study, I figure out how the information might be useful in the real world.
A2Q1Q17 - When I study, I try to understand the material better by relating it to things I already know.
A2Q1W23 - When I study, I figure out how the material fits in with what I have learned.

A2-2 Transformation(alternative)
A2Q1F06 - When I study, I summarise the most important information in my own words.
A2Q1M13 - When I study, I write a short summary of the most important facts.
A2Q1S19 - When I study, I go through my notes and make a diagram showing the most important points.
A2Q1Y25 - When I study, I stop reading to write down the main points of the text.

The two alternative scales behave almost identically. The reliability of the Elaboration scale is between 0.71 and 0.81. The Transformation scale has similar alpha values in most countries. In two countries (Fresia and Gladiola), this scale shows unacceptably low values of alpha: about 0.64.

Some other general remarks on the quality of these two scales:

- The MSP values range from acceptable to good. For the Elaboration scale H values for the 11 countries are found in the range of 0.35 to 0.60, with only one country below 0.40. Values of the Transformation scale are a bit lower, from 0.25 to 0.57, with two counties below the criterion.
- The formulation of some items, for example, A2Q1M13 and A2Q1Y25 as well as A2Q1D04 and A2Q1Q17, is similar and may affects reliabilities in a positive way.

Conclusion: Both scales have good properties. We suggest to include the Elaboration scale, which has an overall better quality than the Transformation scale.

A3: Control strategies

A3Q1AA27 -When I study, I start by figuring out what, exactly, I need to learn.
A3Q1B02 - When I study, I force myself to check to see if I remember what I have learned.
A3Q1I09 - When I study, I try to figure out, as I read, which concepts I still haven't really understood.
A3Q1P16 - When I study, I make sure that I remember the most important things.
A3Q1V22 - When I study, and I don't understand something, I look for additional information To clarify the point.

There is variation among countries on this scale; their mean scores vary highly. For the scale on control strategy, the alpha values are between 0.62 and 0.81. In two countries (Aster and Fresia) the reliability is as low as 0.62. The relevant values for Loevinger's H are between 0.27 and 0.50. Aster and Fresia score below the criterion, namely .27 and .28.

Conclusion on A3: a relatively good candidate.

General conclusion on Learning Strategies:
Instruments for Elaboration and Control Strategies do meet the standards. Memorising does not meet the requirements completely; the consequences are discussed in section 7.

6.3 Motivational preferences

Twenty-items were used to assess motivation.

B1: Compliance (general)
B1Q2K38 - I only study when I have to (Reversed).
B1Q2L39 - I only study when someone makes me (R).

These two items were meant to be used as two separate single items. For reasons mentioned above they are not considered for scale analysis.

Conclusion: will not be applied

B2: Instrumental motivation
B2Q1C03 - I study to increase my job opportunities.
B2Q1J10 - I study to ensure that my future will be financially secure.
B2Q1R18 - I study to get a good job.

There is some variation in the mean value of this scale: it varies to a small extent among countries. The formulation of the first and third items is rather similar, but there is no possibility of deleting one or the other: doing so would cause the scale to disappear. In general, however, the reliability of the instrumental motivation scale is good and is between 0.77 and 0.86 in ten of the eleven countries (in Gladiola it is 0.67). The IRT scaling also shows good properties for most countries. In Gladiola, however, we find the lowest -- but still acceptable -- value of 0.41; for the other countries it is 0.55 or higher.

Conclusion: Instrumental motivation provides a good scale.

B3: Competitive motivation
B3Q1G07 - I study because I want to be one of the best.
B3Q1O15 - I study because I want to show that I'm smarter than others.
B3Q1X24 - I study because I want to perform better than others.

The mean of this scale varies somewhat among countries, indicating relevant perspectives. However, as with B2, items B3Q1O15 and B3Q1X24 are similar, thus inflating the reliability. The alpha values of the scale for competitive motivation is between 0.77 and 0.85 in ten of the eleven countries. In Gladiola it is too low, namely 0.63. However, the IRT scaling values show that we have a good scale, indicated by three values of Loevinger's H around .40; all others are higher then .54.

Conclusion: The scale is relatively good, but has an almost similar content as F2 Competitive Learning. The trade-off will be discussed in section 7.

B4: Interest-based motivation
B4Q1E05 - When I study, I don't notice the time passing.
B4Q1L12 - I study because the subject matter is very important to me.
B4Q1U21 - I study because I am very interested in the subject matter.
B4Q1Z26 - I study because working on the subject matter is fun.

The mean value of this scale varies little between countries, thus making it an unattractive scale. Checking the reliability of the scale with four items produces values between 0.59 and 0.81, but more than half of the countries are below the criterion of 0.70. Excluding the first item improves the scale a little -- reliability varies between 0.66 and 0.84. Nonetheless, in three countries the alpha score is still below 0.70. Without the first item the scale fits well in MSP: between 0.49 and 0.61.

Conclusion: The scale does not meet minimal requirements for reliability. It overlaps considerably with the subject-specific scales of B7.

B5: Compliance (subject related)
B5Q2M40 - I only read when I have to (R).
B5Q2N41 - I only do math when I have to (R).

These two items relate to two different domains and thus will not be elaborated further (no single item indicators).

Conclusion: will not be applied

B6: Instrumental motivation (subject related)
B6Q2C30 - To get ahead, it is important for me to read.
B6Q2H35 - To get ahead, it is important for me to know math.

These two items relate to two different domains and thus will not be elaborated further (no single item indicators).

Conclusion: will not be applied

B7: Interest (subject related)
Math
B7Q2B29 - I do math in my spare time.
B7Q2E32 - When I do math, I sometimes get totally absorbed.
B7Q2G34 - Math is important to me personally.
B7Q2J37 - Because doing math is fun, I wouldn't want to give it up.
Verbal
B7Q2A28 - Reading is important to me personally.
B7Q2D31 - Because reading is fun, I wouldn't want to give it up.
B7Q2F33 - I read in my spare time.
B7Q2I36 - When I read, I sometimes get totally absorbed.

The mathematics scale does vary between countries. The verbal scale varies to a small extent. Boys score higher than girls on the mathematics scale (except in Gladiola) and lower on the verbal scale than girls in all countries. Nevertheless, the formulation of items raises some questions. “I read in my spare time” is a question of relevance to most students. In contrast, ”Doing math in the spare time” is quite uncommon. The average on a scale from 1 to 4 is 1.9 (SD: 1.0). The range of this item between countries is 1.6 to 2.4. It is probably unwise to have these corresponding questions. After excluding the “dubious” math item, the mathematics motivation scale is highly reliable. Alpha values for the mathematics motivation scale range between 0.71 and 0.90. The three-item scale fits very well in MSP.

The scale for verbal motivation is also of high quality and shows reliability values between 0.78 and 0.90. B7Q2A28 is the item that is most appropriate to remove in the verbal scale. Excluding that item, the verbal motivation scale is still highly reliable. The three-item scale also fits very well in MSP.

Conclusions on Subject related interest: Two high-quality scale have been detected, interest in math (three items) and interest in verbal (three items). Since these scales have better properties than B4, we prefer to include these B7m and B7v.

General conclusions on Motivational Preferences:
Three high-quality instruments in this dimension can be used: Instrumental Motivation (B2), Math Interest (B7m), and Verbal Interest (B7v). The other instruments do not meet the standards set for the comparative PISA measurement. [7]

[7] We also tried to extend B7 with B6. In some countries, (e.g., Jasmine), B2, B3, and B4 correlate highly (0.56-0.74), but in Fresia the correlations are low (0.17-0.26). Another option was to combine B1 and B5. For three countries: Daffodil (α = 0.80; mean = 8.5; SD = 3.0), Fresia (α = 0.76; mean = 10.7; SD = 3.4), and Lily (α = 0.73; mean = 10.1; SD = 3.1). This did not lead to the desired reduction of items with increased quality.

6.4 Goal Orientation

Goal Orientations is defined according to two different subdimensions: task orientation and ego orientation with nine items.

C1: Task orientation
C1Q3B02 - I feel most successful if what I have learned really makes sense to me.
C1Q3G07 - I feel most successful if I discover a new approach to solving a problem.
C1Q3L12 - I feel most successful if a class makes me think about things.
C1Q3Q17 - I feel most successful if something I studied makes me want to find out more.
C1Q3T20 - I feel most successful if I finally master a complex problem.

The empirical findings on this scale were not particularly satisfactory. The mean of this scale varies only to a small extent among countries, thus making it a less interesting candidate. In addition, the reliability of the task orientation scale is low and ranges between 0.62 and 0.78. In more than half the countries (Anemone, Aster, Carnation, Fresia, Gladiola, and Jasmine), alpha values are below 0.70. The scalability check in MSP showed a similar picture: most values are well below 0.40 and only a few are above the criterion. Taking these bad results from MSP into account, we did not attempt to use LISREL to assess the model's fit.

Conclusion: Task Orientation is not a good scale

C2: Ego orientation
C2Q3D04 - I feel most successful if I can demonstrate that I'm smart.
C2Q3I09 - I feel most successful if I get better grades than the others.
C2Q3O15 - I feel most successful if I'm the only one who knows the right answer.
C2Q3U21 - I feel most successful if I know more than the others.

This appears to be a good scale. The mean of this scale varies among countries. Boys score higher on this orientation than girls in most countries, which makes further research relevant. Also the reliability of the ego orientation scale is acceptable and ranges from 0.75 and 0.85 in ten of the eleven countries. In the other country, Carnation, the alpha value is 0.67. In the IRT analysis with MSP the scale does fit well: all values exceed 0.40 up to 0.63. As a check, we fitted the scale in LISREL with good results.[8]

Conclusion: Ego Orientation is a good scale

General conclusion on Goal Orientation

Goal orientation should be represented only by ego orientation. The conceptual basis for using two distinct scales cannot be justified analytically.[9] Having two scales for goal

[8] For Fresia: factor loadings: 0.37 and higher; $\chi^2 = 2.5$ (df=2) p=0.29; root mean square error of approximation (RMSEA) = 0.025; Standardized RMR (root mean square residual) = 0.016; Non-normed fit index (NNFI) = 1.00; For Jasmine: factor loadings: 0.51 and higher; $\chi^2 = 1.0$ (df=2) p=0.59; root mean square error of approximation (RMSEA) = 0.0; Standardized RMR (root mean square residual) = 0.013; Non-normed fit index (NNFI) = 1.02.

[9] A scale for goal orientation with nine items (C1 plus C2) had reliability values between 0.73 and 0.86. MSP values were between 0.27 and 0.46. Most problems appear to be caused by C1Q3B02 and C1Q3L12. The content of C1 is

orientation would cause an unbalanced measure of the dimension, which will be discussed in Section 7.

similar to B4 and B7 (interest-based motivation); C2 is similar to B2 and B3 (competitive/instrumental motivation). Due to the different packets we could not empirically evaluate the overlap.

6.5 Self-Related Cognitions

The area of self-related cognitions was approached with a series of six instruments, comprising 33 items. Some of the instruments were tested as alternatives for each other, so a reduction of the number of scales is feasible.

D1: Agency beliefs: effort
D1Q4A23 - I can really pay attention when I'm trying to learn something.
D1Q4B24 - It's not hard for me to really put in enough effort in school.
D1Q4D26 - When it comes down to learning, I can really work hard.
D1Q4F28 - Making myself listen carefully to my teachers is easy.
D1Q4H30 - If I decide to, I can listen very carefully.

The mean of this scale varies among countries. No meaningful differences between boys and girls were found. Reliability of this agency belief scale is good and lies between 0.70 and 0.84 in most the countries. Only in Fresia is reliability on the low side (i.e., 0.67). After excluding a disturbing item, 24, alpha continues to vary between 0.70 and 0.81. Fresia has a low value of 0.63.

The scalability of the items is satisfactory: the MSP values range from 0.34 and 0.59. When item D1Q4B24 was removed, these values remained the same. A test of the scale in LISREL showed satisfactory results, indicating homogenity.[10]

Conclusion: After the removal of item B24 a good four-item scale for Agency Beliefs (Effort) is available.

D2-1 Agency beliefs: ability
D2Q4C25 - I can learn the things I need to learn pretty fast, without really trying a lot.
D2Q4E27 - I'm pretty smart-even without working very hard.
D2Q4G29 - When it comes to learning, I'm pretty smart.

Reliability of this second agency belief scale is less good and lies between 0.58 and 0.84 in most the countries. In three countries (Cornflower, Fresia and Gladiola) reliability is below 0.70. The scalability of the items also varies largely: the MSP values range from 0.28 to 0.81. Since we have no opportunity to delete bad fitting items to improve the scale, the overall judgement must be: moderate scale.

Conclusion: A moderate scale for Agency beliefs: ability can be developed.

D2-2: Control expectation (alternative)
D2Q4I31 - When I sit myself down to learn something really hard, I can learn it.
D2Q4J32 - If I decide not to get any bad grades, I can really do it.
D2Q4K33 - If I decide not to get any problems wrong, I can really do it.
D2Q4L34 - If I want to learn something well, I can.

[10] For Jasmine: factor loadings: 0.51 and higher; $\chi^2 = 8.4$ (df=5) p=0.13; root mean square error of approximation (RMSEA) = 0.054; Standardized RMR (root mean square residual) = 0.020; Non-normed fit index (NNFI) = 0.99

This is a good scale. The reliability estimates range from 0.69 to 0.84.
MSP values are all satisfactory and range from 0.40 to 0.74 No specific item could be deleted to improve the scale even further.

Conclusion: A good scale for control expectation can be developed.

D2-3: Self-efficacy (alternative)	
D2Q4M35 -	I believe I will receive excellent grades.
D2Q4N36 -	I'm certain I can understand the most difficult material presented in readings
D2Q4O37 -	I'm confident I can understand the basic concepts taught.
D2Q4P38 -	I'm confident I can understand the most complex material presented by the teacher.
D2Q4Q39 -	I'm confident I can do an excellent job on assignments and tests.
D2Q4R40 -	When studying, I expect to do well.
D2Q4S41 -	I'm certain I can master the skills being taught.
D2Q4T42 -	Considering the difficulty of the subject matter, the teacher, and my skills, I think I will do well.

Since it was also the purpose of the analysis to reduce the total number of items, we considered the overall properties of the scales.[11] This scale (with eight items) has good prospects for further analysis. Using LISREL, we reduced the eight-item self-efficacy scale to four items: D2Q4N36, D2Q4P38, D2Q4Q39, and D2Q4S41.[12] This scale matched the IRT criteria very well (range 0.41 to 0.68) and had lower boundaries of reliability of 0.78, again with the exception of Fresia (0.67).

Conclusion: A good scale for self-efficacy can be developed.

Intermediate conclusion: The two scales on Control Expectation and Self Efficacy were seen as alternatives for two Agency Beliefs (Effort and Ability) scales. Over the whole range the first mentioned pair of scales (good, good) show better properties then the second ones (good, moderate).
Thus we prefer to apply the scales for Control Expectation and Self Efficacy.

D3: Self-concept verbal	
D3Q3A01 -	I'm hopeless in English classes (R).
D3Q3J10 -	I learn things quickly in English classes.
D3Q3R18 -	I get good marks in English.

[11] Correlation between D1 and D2-2 is 0.61; between D2-1 and D2-3 is 0.69.

[12] Results for Snowdrop: factor loadings: 0.47 and higher; $\chi^2 = 3.1$ (df=2) p=0.21; root mean square error of approximation (RMSEA) = 0.030; Standardized RMR (root mean square residual) = 0.012; Non-normed fit index (NNFI) = 1.00.
Results for Jasmine: factor loadings: 0.78 and higher; $\chi^2 = 0.9$ (df=2) p=0.64; root mean square error of approximation (RMSEA) = 0.0; Standardized RMR (root mean square residual) = 0.0063; Non-normed fit index (NNFI) = 1.01
Results for Fresia: factor loadings: 0.34 and higher; $\chi^2 = 0.7$ (df=2) p=0.72; root mean square error of approximation (RMSEA) = 0.0; Standardized RMR (root mean square residual) = 0.0100; Non-normed fit index (NNFI) = 1.04

This appears to be an excellent scale. The mean varies to a small extent among countries. Boys score lower than girls in most countries. The reliability of the scale meets the standard, with values between 0.75 and 0.84. The IRT values in MSP are 0.56 or higher.

Conclusion: Self-Concept Verbal is an excellent scale.

D4: Self-concept math
D4Q3F06 - I get good marks in mathematics.
D4Q3K11 - Mathematics is one of my best subjects.
D4Q3P16 - I have always done well in mathematics.

This appears to be an excellent scale. The mean of this scale varies among countries. Boys score higher than girls do in most countries. The reliability of the scale varies and always meets the standard, with values above 0.84. The IRT values in MSP are 0.65 or higher.

Conclusion: Self-Concept Math is an excellent scale.

D5: Self-concept academic
D5Q3E05 - I learn things quickly in most school subjects.
D5Q3N14 - I do well in tests in most school subjects.
D5Q3V22 - I'm good at most school subjects.

This appears to be a good scale. The mean of this scale varies among countries. No differences are noticed between boys and girls in most countries. The reliability scores meet the critical value in all countries and vary between 0.76 and 0.84. The values in MSP of 0.56 or higher confirm the scale's merits.

Conclusion: Self Concept Academic is a very good scale.

D6: Self-concept general
D6Q3C03 - Overall, I have a lot to be proud of.
D6Q3H08 - Overall, I'm a failure (R).
D6Q3M13 - Most things I do, I do well.
D6Q3S19 - If I really try, I can do almost anything I want to do.

The descriptives show that mean of this scale varies among countries, but the scale's reliability values are lower than 0.70 in all countries. This is not reassuring, as is supported by the MSP analysis. In addition, the values are generally below the acceptable threshold. With the poor results from MSP in mind, we did not try to fit the scale in LISREL.

Conclusion: No good scale can be developed with these items.

General conclusion on Self-Related Cognitions
This dimension can be represented by five scales with a total of 17 items. The scales are Control Expectation (D2-2), Self Efficacy (D2-3), Self-Concept Verbal (D3), Self-Concept Math (D4), and Academic Self Concept (D5). The analysis confirmed the imprudence of combining the D3, D4 and D5 scales into one instrument.[13]

[13] D3 and D4 are not correlated. The correlation between D5, and D3 or D4 is about 0.50.

6.6 Action Control: Effort and Persistence

Two scales with 10 items were used to measure action control. One scale emphasises the general aspects of effort and persistence in learning; the other focuses on subject-related issues.

E1: Effort and Persistence in learning: general
E1Q5C03 - When studying, I work as hard as possible.
E1Q5E05 - When studying, I keep working even if the material is difficult.
E1Q5G07 - When studying, I try to do my best to acquire the knowledge and skills taught.
E1Q5I09 - When studying, I put forth my best effort.

This is an excellent scale. The mean of this scale varies to a small extent among countries, and there is also variation according to gender. Boys score lower on effort and persistence than girls in all countries. The reliability of the scale (general) varies between 0.76 and 0.87. The IRT analysis with MSP also demonstrates the scale's good properties (range from 0.49 to 0.69). This scale also fits well in LISREL.[14]

Conclusions: An excellent scale

E2: Effort and Persistence: subject related
Math
E2Q5B02 - I am ambitious in trying to achieve good grades in math.
E2Q5F06 - I work hard in my math class
E2Q5J10 - I persist when I have to cope with a math problem.
Verbal
E2Q5A01 - I work hard in my English class.
E2Q5D04 - I persist when I have to cope with a difficult or long text.
E2Q5H08 - I am ambitious in trying to achieve good grades in English.

The reliability of the mathematics part of the effort and persistence scale with three items varies between 0.59 and 0.87. In two countries (Cornflower and Snowdrop), alpha values are below 0.70. In MSP the scale has values of 0.46 or higher., but with a dramatic low value of 0.17 for Snowdrop.

(footnote 13 cintinued)
LISREL results for D3, D4 and D5:
For Jasmine: factor loadings: 0.62 and higher; χ^2 = 78.0 (df=24) p=0.00; root mean square error of approximation (RMSEA) = 0.097; Standardized RMR (root mean square residual) = 0.063; Non-normed fit index (NNFI) = 0.88
For Fresia: factor loadings: 0.46 and higher; χ^2 = 58.0 (df=24) p=0.00; root mean square error of approximation (RMSEA) = 0.063; Standardized RMR (root mean square residual) = 0.046; Non-normed fit index (NNFI) = 0.95;
Modification indices indicate that:
a) The error terms of some items of D3, D4, and D5 may be correlated.
b) Some items of those scales have a high factor loading on other scales.
This is not reassuring.

[14] For Jasmine: factor loadings: 0.84 and higher; χ^2 = 4.1 (df=2) p=0.13; root mean square error of approximation (RMSEA) = 0.065; Standardized RMR (root mean square residual) = 0.0098; Non-normed fit index (NNFI) = 0.99

The reliability of the verbal part of the effort and persistence scale with three items is much lower and varies between 0.59 and 0.77. In five countries (Aster, Cornflower, Fresia, Jasmine, and Snowdrop) alpha values are below 0.70. This makes the scale an unattractive instrument, also since the MSP analysis suggests moderate scaling.

Conclusion: The two scales on subject-related effort and persistence do not demonstrate sufficient quality to be pairwise included in a final selection.

General conclusion on Action Control: Effort and Persistence
A general and a subject-specific component have been distinguished. The general scale (E1) is an excellent scale that should be applied. In contrast, the pair subject-specific scales (E2) hardly not meet the standard and should not be included.[15]

[15] One might argue that an analysis on the 10 items from both scales would provide evidence for a scale that includes both general and subject-specific information. There are two arguments against this. First, our goal is to reduce the number of items, not to retain the present number. Second, empirical arguments also count. Though reliabilities rise for the total scale (a function of number of items), the MSP parameters show a **decrease** of the values. This suggests that it is not wise to merge the scales.

6.7 Learning Preferences

Learning preferences belong to the core of the Self-Regulated Learning competencies. Typically, two aspects have been distinguished: cooperation and competition. measured with eight items.

F1: Cooperative learning
F1Q6A11 - Working in a group now helps me work with other people later.
F1Q6D14 - I do not like working with other people (R).
F1Q6G17 - Working in a group scares me (R).
F1Q6J20 - We get the work done faster if we all work together.

This scale does not meet the expectations. The items appear to be very heterogeneous (anxiety to cooperate, advantages of cooperating, etc.) and the mean values vary a small extent among countries. The reliability of the cooperative learning scale is 0.65 or lower, a result confirmed by the IRT analysis. The scale does not fit in MSP.

Conclusion: Not a suitable scale.

F2: Competitive learning
F2Q6C13 - I like to try to be better than other students.
F2Q6F16 - Trying to be better than others makes me work well.
F2Q6H18 - I would like to be the best at something.
F2Q6K21 - I learn faster if I'm trying to do better than the others.

The mean of this scale varies to a small extent among countries. Boys scored higher than girls in all countries. The scale's reliability is between 0.74 and 0.81 in ten of the eleven countries. Only Gladiola produces an outlier, with an alpha value of 0.56. Since this country also had the highest average scores, this may be caused by restriction of range. The MSP values of 0.40 and higher suggest that we have a relative good scale (with two low exceptions of Aster and Gladiola). This scale also fits well in LISREL.[16]

Conclusion: a relative good scale to measure competitive learning[17]

General conclusion: Only a good scale for competitive learning (F2) is available to represent this dimension. If this is the only indicator for this domain, then there is a lack of balance. This will be discussed in section 7.

---

[16] For Jasmine: factor loadings: 0.71 and higher; $\chi^2 = 3.0$ (df=2) p=0.23; root mean square error of approximation (RMSEA) = 0.043; Standardized RMR (root mean square residual) = 0.012; Non-normed fit index (NNFI) = 1.00
For Fresia: factor loadings: 0.39 and higher; $\chi^2 = 1.4$ (df=2) p=0.50; root mean square error of approximation (RMSEA) = 0.0; Standardized RMR (root mean square residual) = 0.014; Non-normed fit index (NNFI) = 1.01.

[17] The content of F2 is similar to B3. F2 is more general than B3, and the questions are better worded.

6.8 Implicit Theories of Learning – stability of learning potential

In the conceptual thinking on self-regulated learning instruments were used to measure more implicit attitudes relevant to the learning process. Three scales were candidates in this respect: opinions about natural ability and on the relevance of effort and ability for performance. The analysis allows us to determine whether these attitudes can be measured satisfactorily.

G1: stability of learning potential
G1Q6B12 - Natural ability determines how fast one can learn (R).
G1Q6E15 - There are natural individual differences in learning potential (R).
G1Q6I19 - No matter how hard some students try, they will never become good learners (R).

The reliability for this scale is generally 0.60 or lower. This scale does not fit in MSP, which makes it unattractive for comparative applications.

Conclusion: Not a good scale.

G2: Importance of Effort for Performance
G2Q6L22 - I believe that studying hard is the main factor in learning.
G2Q6M23 - I believe that effort is the main factor determining my academic performance.
G2Q6N24 - I feel that my academic performance is determined by my effort.
G2Q6O25 - I believe that if I fail, it is because I do not study hard enough.

Items related to the importance of effort do not constitute a good scale. The scale's reliability varied between 0.61 and 0.74. In at least five countries, reliability was below 0.70. When the last item (G2Q6O25) was deleted, reliability increased somewhat in most countries but did not solve the reliability issue. Scalability according to MSP was not sufficient: half the countries scored lower then 0.40. Consequently, an alternative strategy was applied. The content of the G2 scale is somewhat similar to E1 (effort and persistence). The correlation between E1 and G2 averages about 0.30, but adding the G2 scale to the E1 scale did not result in a better scale for effort and persistence.

Conclusion: Not a good scale.

G3: Importance of Ability for Performance
G3Q6P26 - Intelligence cannot be changed (R).
G3Q6Q27 - The smarter you are, the less time you need to study (R).
G3Q6R28 - Geniuses are born, not made (R).
G3Q6S29 - You don't need to try very hard to get good grades if you are smart (R).

The counterpart of G2 (effort) is in this scale G3 (ability). Consequently, it is only relevant to have the pair of scales and not just one. Unfortunately, the reliability of this implicit theory of learning scale is on the low side. Alpha values are between 0.60 and 0.74 and in most countries are below 0.70. Excluding the first item (G3Q6P26) increased the reliability somewhat in most countries, but not enough. Use of MSP revealed that the scalability was below standard.

Conclusion: Not a good scale.

General Conclusion: Implicit Learning Theories cannot be measured satisfactorily with the three proposed scales.

7 Conclusions of the Analysis

The core issue in the analysis of the data of the CCC field trial has been the reduction of the number of items for the scales. Scales and/or dimensions will be deleted or re-arranged primarily on the basis of two criteria: theoretical content arguments and statistical quality.

Selecting the Scales

In most instances these criteria could be applied properly and produce meaningful results that allow judgement about the feasibility and appropriateness of including or discarding items or scales. Despite this success, two issues remained unsolved.

First, overlap between items should be avoided. If all items had been in the each packet, the statistical analysis could address the problem of overlap by identifying redundant items. As is the case here, however, when similar items or scales were included in different packets and administered to different sets of students, no statistical analysis can be applied. This overlap between packets happens to be the case with Competitive Motivation (B3) and Competitive Learning (F2).

Second, three decisions on the acceptability of a scale were postponed because they required a content rather than a statistical justification.

Content decision 1: Memorising:

The scale has weaker properties than needed for comparative purposes. Nevertheless, the important dimension on Learning Strategies cannot be justified theoretically if a conceptually relevant aspect is omitted. With the experts the research team agrees that the scale on memorising can be applied, although on an experimental basis. The scale should be flagged and its statistical properties established again after the PISA data are collected in 2000.

Content decision 2: Goal Orientation

Two scales, task and ego orientation represent goal orientation. The statistical properties of the task orientation scale are too low to justify its use. In contrast, the scale for ego orientation is of good quality. Unfortunately, with only one of the two scales meeting the statistical requirements, an effort to measure goal orientation with a single scale would be unbalanced. There is no alternative available, so goal orientation should not be assessed in PISA.[18]

Content decision 3: Cooperative learning

Two scales were used to assess learning preferences: cooperative (F1) and competitive learning (F2). The latter scale had a high quality, but this cannot be said about the scale for cooperative learning. Cooperation is an important dimension of self-regulated learning[19]. Omission of items related to cooperation would put an unjustifiably strong emphasis on the competitive character of education. To address this concern we recommend that the presently

[18] Also, the content of the Ego scale is rather similar to that of the F2 Competitive Learning preference. See the earlier discussion of Marsh et al (1999) showing these to be highly correlated (r=.90).

[19] This also justifies the trade-off between B3 and F2 in favor of F2. Otherwise the complete domain of Learning Strategies would have disappeared.

unsatisfactory items in scale F1 items be replaced with four items that have proven to be of high quality in other studies of cooperative learning (Marsh et al, 1999). With the experts we agree that this scale[20] on cooperation can be applied, though on an experimental basis. The scale should be flagged and its statistical properties established again after the PISA data are collected in 2000.

The Time Constraint

One of the purposes of the present analysis was to reduce the number of items in such a way that the resulting set can be administered in no more than 10 minutes. The selection made above consists of 51 items. In the field trial and in preceding test stages, the modal time for the completion of 40 items was eight minutes, or five items per minute. If we apply this as a yardstick for our present selection, the available test time of 10 minutes will not be exceeded.

[20] The scale has the following items, with reported reliabilities well above .80:
- I like to work with other students
- I learn the most when I work with other students
- I do my best work when I work with other students
- I like to help other people do well in a group
- (it is helpful to put together everyone's ideas when working on a project)

Table 4: Overview of the results of the analyses

Dimension	*Name*	Code l	*# Start*	*# Final*	*Overall quality*	*Decision/ Remarks*
A: Self Regulated Learning	*Memorising*	**A1**	4	?4?	Not strong	Content decision 1
	Elaboration	**A2-1**	4	4	Good	Apply
	Transformation	**A2**	4	-	Good	Favor A2-1, Not apply
	Control Strategies	**A3**	5	5	Good	Apply
B: Motivational Preferences	*Compliance: Gen*	**B1**	2 x 1	-	Not analysed	Not apply
	Instrumental M	**B2**	3	3	Good	Apply
	Competitive M	**B3**	3	-	Good	Favor F2, Not apply
	Interest based M	**B4**	4	-	Not strong	Not apply
	Compliance: Spec	**B5**	2 x 1	-	Not analysed	Not apply
	Instrumental M	**B6**	2 x 1	-	Not analysed	Not apply
	Interest *Mathematical* *Verbal*	 **B7-m** **B7-v**	 4 4	 3 3	 Very good Very good	 Apply Apply
C Goal Orientation	*Task Orientation*	**C1**	5	?	Not good	Content decision 2
	Ego Orientation	**C2**	4	?	Good	
D Self Related Cognitions	*Agency Beliefs: Effort*	**D1**	5	-	Good	Favor D2-2. Not apply
	Agency Beliefs: Ability	**D2-1**	3	-	Moderate	Favor D2-3, Not apply
	Control Expectation	**D2-2**	4	4	Good	Apply
	Self Efficacy	**D2-3**	8	4	Good	Apply
	Self Concept Verbal	**D3**	3	3	Excellent	Apply
	Self Concept Math	**D4**	3	3	Excellent	Apply
	Self Concept Academic	**D5**	3	3	Very good	Apply
	Self Concept General	**D6**	4	-	Not good	Not apply
E Action Control: Effort and Persistence	*Effort and Persistence: General*	**E1**	4	4	Excellent	Apply
	Effort and Persistence: Subject	**E2**	2 x 3	-	Moderate	Not apply
F Learning Preferences	*Cooperative Learning*	**F1**	4	?4?	Not good	Content decision 3
	Competitive Learning	**F2**	4	4	Relatively good	Apply
G Implicit Theories of Learning	*Stability*	**G1**	3	-	Not good	Not apply
	Effort	**G2**	4	-	Not good	Not apply
	Ability	**G3**	4	-	Not good	Not apply
		Total	112	43/ 8?		

8 Quality of the Scales for PISA

8.1 Reliability of the Scales

Given the relative complexity of the scales, the importance of international comparisons, and the high levels of attention that will be directed at PISA's results, it is imperative that the dimensions of self-regulated learning reflect robust levels of reliability. For this reason we have established the reliability of each of the scales for each country and for each subpopulation to be examined in each country.

Reliability has been expressed as a value of Cronbach's alpha for each country and for each subscale. In addition, we used the slightly larger flexibility of IRT in two ways. First, some IRT models have their own reliability index with a similar interpretation: the index of subject separation (basically the reliability from classical test theory assessed on the latent trait axis) or the Mokken-Molenaar-Sijtsma reliability estimate based on interpolation using nonintersecting item response functions. Second, IRT offered the estimated test information function as an expression for the measurement precision for each separate latent trait value.

In Table 5 an overview is given of the proportion of scales that meet the reliability criterion of 0.70. For the total sample of 22 countries this is 88%. The data from the second wave seem to produce somewhat lower reliabilities. The exact reason is yet unkown. From the available descripitive information, however, it is suggested that a ceiling effect may be the cause, in particular for countries like Honeysuckle and Violet. Sometimes also the small number of cases in a country in particular in a breakdown to gender or SES, causes low reliabilities. In the PISA study with larger samples this effect is likely to disappear.

Table 5: Overview of proportion of high reliability (Chronbachs' alpha >= 0.70)
(SC is # scales x # countries)

	1st Wave 11 countries	2nd Wave 11 countries	Total 22 countries
	(SC=132)	(SC=167)	(SC=299)
Total sample	0,95	0,82	0,88
	(SC=132)	(SC=167)	(SC=299)
Boys	0,92	0,84	0,87
Girls	0,93	0,77	0,84
	(SC=120)	(SC=75)	(SC=195)
Low SES	0,76	0,68	0,73
Medium SES	0,94	0,80	0,89
High SES	0,89	0,83	0,87

These outcomes confirmed that concepts have been reliably measured. Our strategy of converging evidence assured that the decisions taken are based on different research methods.

This is the more important in those cases where the alpha value raises doubts: it facilitated a judgement on the nature of the problem and the best way to address it.

In addition we have approached the reliability issue in two other ways.

First, we rely on the NPMs monitoring of the data collection as a source information on the "effective" functioning of items and translation issues. We have asked the NPMs to alert us to potential problems (see section 5.3). This is particularly relevant since in some cultural areas the reliability (and scalability) coefficients seem to be on the low side. This applies in particular to Fresia, Gladiola, Honeysuckle, Violet and to a lesser extent to Rose.

Second, the expert panel took a vital role in these matters of judgement regarding issues of reliability and of validity.

8.2 Construct Validity

We make a distinction between two types of construct validity, namely the *validation of a measurement instrument* and the *validation of a theory*. For the validation of the instruments of self-regulated learning we rely heavily on the structure of the scales as discussed by the authors of the conceptual paper (Baumert *et al*, 1998). On the basis of the confirmatory analysis, it is clear in which cases the anticipated structure has been found in the empirical data. Thus, the CFA and IRT analyses on the separate scales already have provided support for the construct validity of the instruments. Our expectations with regard to logistic IRT models, however, could not be met[21].

No specific theory on self-regulated learning has been formulated and hence the analysis did not try to test such. Nevertheless, the research literature and the conceptual work done earlier provides us with expectations that can be used for additional validation. Most scales are expected to have positive relationships with the measures of student achievement. If they fail to do so -- or if they have a negative relationship -- for one or more countries, this cannot be seen as contributing to construct validity. The correlations to be found in the field trial should at least resemble those reported in the literature and show signs of relationships in the anticipated directions.

Table 6 provides the relevant correlations for the selected scales and supports the validity of the Self Regulated Learning scales in this respect

[21] The scaling of attitudes with -- in our case -- a limited number of items each presented with four ordered answer categories frequently leads to disappointing results. This holds a fortiori not only if the frequencies per item category are fitted according to a parametric IRT model, but it is also required that the item-category location parameters are invariant across - in our case - 11 countries, two gender groups, and three SES groups.
Our main attempts were based on the OPLM software package (Verhelst et al., 1995). Some were cross validated using ConQuest (Wu et al., 1998) with disappointing results. From detailed output for the scales A1, B7v, B7m, E1 and F2, we find that all goodness-of-fit tests from OPLM were dramatically significant. Moreover, none of the usual strategies seem to work: the misfit is not located in one specific item, one specific category, or in one specific country. In addition, the OPLM feature of postulating unequal integer slopes per item, which keeps the CML estimation alive (Verhelst et al., 1995), did not solve our problem. The nonparametric package MSP was designed for the scaling of attitudes (ConQuest and OPLM have proven most efficient in scaling achievement tests). Given that MSP is useful in finding appropriate subscales and assessing their quality, it was thus decided, at least for now, not to pursue parametric IRT models for the CCC data.

Table 6: For Construct Validity Only: Tentative Correlations between Self-Regulated Learning scales and Achievement Tests (N=11 countries, separate packets, NOT-representative samples)

Dimension	Scale	Label	# Items	Mathematics	Reading	Science
Self-Regulated Learning	*Memorising**	*A1*	*4*	*-0.07*	-	*-0.06*
	Elaboration	A2-1	4	0.12	0.10	0.13
	Control Strategies	A3	5	0.11	0.17	0.13
Motivational preferences	Instrumental Motivation	B2	3	-	0.04	-
	Interest math	B7 math	4	0.19	-	0.06
	Interest verbal	B7 verb	4	0.17	0.28	0.23
Self Related Cognitions	Control Expectation	D2-2	4	0.21	0.19	0.21
	Self-Efficacy	D2-3	4	0.30	0.21	0.26
	Self-Concept Verbal	D3	3	0.13	0.25	0.19
	Self-Concept Math	D4	3	0.44	0.22	0.26
	Self-Concept Academic	D5	3	0.33	0.30	0.32
Action Control: Effort and Persistence	Effort and Persistence in Learning: General	E1	4		0.16	
Learning Preferences	*Cooperative **	*F1*	*4*		*0.08*	
	Competitive L	F2	4		0.08	

* These scales are suggested to remain included on the basis of conceptual coverage of the domain.

Determination of acceptable levels of validity are important not only at the national level but also for each subpopulation (gender, SES) that will be the subject of analysis within each country. This is typically a time-consuming process and will require restriction to a limited number of subpopulations. We have focussed on SES groups and gender. Interactions of gender and SES with achievement have not been further analyzed, because no differential effects were found. The analyses were replicated for each participating country.

These issues are sensitive and have been addressed with care to prevent bias in the scales both within and between the countries. The panel of experts devoted special attention to this issue and we will inform the NPMs of any major anomalies or unanticipated discrepancies.

To get an impression of the relevance of these coefficients when background characteristics are taken into account, we also calculated the correlations between the proposed scales and the achievement scores in LISREL. In Table 7 it becomes clear that that background variables (SES, gender) indeed play an intermediate role. Substantial effects become visible. Mathematics

achievement seems to be effected most by Self Concept Math (D4) and to a lesser extent the interest in this subject (B7 m) and the Academic Self Concept (D5). Verbal Self Concept (b7v) and Academic Self Concept (D5) show the strongest effects on reading achievement. And for Science apparently Verbal Interest (B7v) and Academic Self Concept (D5) play an important role. For the moment no further analyses can be done. We recall that the packets contained different scales and thus the relative importance cannot be established until the data collection of the PISA Main Study in 2000.

Table 7 **For Construct Validity Purposes ONLY: Tentative Beta weights (standardised regresseion coefficients), controlling for background variables (N=11 countries, separate packets, NOT-representative samples)**

Dimension	Scale	Label	# Items	Mathematics	Reading	Science
Self-Regulated Learning	*Memorising**	*A1*	*4*	.	-	.
	Elaboration	A2-1	4	.12	.09	.11
	Control Strategies	A3	5	-.02	.09	.07
Motivational preferences	Instrumental Motivation	B2	3	-.06	-.02	-.07
	Interest math	B7 math	4	.14	-.03	-.01
	Interest verbal	B7 verb	4	.16	.26	.29
Self Related Cognitions	Control Expectation	D2-2	4	-.10	-.04	-.09
	Self-Efficacy	D2-3	4	.13	.07	.15
	Self-Concept Verbal	D3	3	.04	.09	-.02
	Self-Concept Math	D4	3	.39	.10	.16
	Self-Concept Academic	D5	3	.16	.24	.22
Action Control: Effort and Persistence	Effort and Persistence in Learning: General	E1	4	-	.11	-
Learning Preferences	*Cooperative **	*F1*	*4*	.	.	.
	Competitive L	F2	4	-	.10	-

* These scales are suggested to remain included on the basis of conceptual coverage of the domain.

In addition it appears that Student Job Expectation (which is also available in the PISA Questionnaire) correlates positively with all scales. Effects of gender are all in the expected directions and match the findings reported in the analyses of the scales.[22]

[22] The factor loadings of all the items of the twelve proposed scales are above 0,52.

8.3 Related Issues

One of PISA's major objectives is comparison among nations. For such comparisons to be valid, the survey items must have equivalent meaning in each country. This is more than an issue of translation from one language to another (which has to be secured by the translation protocols of PISA). With more than 20 countries participating in the assessment of self-regulated learning, ensuring equivalence and comparable levels of acceptable reliability and validity may be problematic.

In a study like this some issues always remain untreated or unsolved. Nevertheless, we think that most issues have been solved. A special check has been made to verify our position that simple unweighted sums of score are adequate in the CCC study. It has been investigated how much gain could be obtained by not reporting the simple sum score across items and categories. Alternatives that come to mind are weighted sums, or translations to a latent trait scale.

As regards the former, it was found that in the 11 countries of the first wave the simple sum score correlated more than .99 with the score using the item weights suggested by OPLM. [23] It is important, in the present application of IRT, that the results remain as transparent as possible to non-IRT specialists. Given the tiny differences and the major loss of transparency resulting from the use of more complicated scalings, it is recommended to stick to the simple unweighted sum score. [24]

We have also undertaken separate LISREL analyses: they are reported in the Technical Report in more detail. At this place it is enough to say that our decisions on scale selection were not effected by a choice for any specific estimation method.

An important last criterion has been whether there is enough variation in the scores on the CCC instruments to raise scientific and policy interest. With the discussion on each scale we have given relevant information, which is based on the information in the Technical Report.

[23] . As regards the latter, both the unweighted and the weighted version were found to correlate .98 or more with the estimated theta values (Warm-corrected, with CML item category parameters inserted).

[24] It could also be linearly translated to a suitably chosen mean and variance, but even there it is doubtful whether the gain is larger than the loss.

9 Presentation of the scales

On the basis of the foregoing we can now present the complete list of scales and items to be applied in the PISA Main Study indicating self-regulated learning as a cross-curricular competency. On the basis of the remaining instruments we suggest to re-group these into three new overarching concepts: Learning Strategies, Motivation and Self-Concept. The selected scales have been assigned to their new position in the conceptual framework and all items are listed as in the order as they should appear in the Main Study Questionnaire.

SELF REGULATED LEARNING QUESTIONNAIRE

			almost never	some-times	often	almost always
1	MEM-1	When I study, I try to memorise everything that might be covered.	☐1	☐2	☐3	☐4
2	SE-1	I'm certain I can understand the most difficult material presented in readings	☐1	☐2	☐3	☐4
3	CS-1	When I study, I start by figuring out what exactly I need to learn.	☐1	☐2	☐3	☐4
4	CE-1	When I sit myself down to learn something really hard, I can learn it.	☐1	☐2	☐3	☐4
5	MEM-2	When I study, I memorise as much as possible	☐1	☐2	☐3	☐4
6	IMOT-1	I study to increase my job opportunities	☐1	☐2	☐3	☐4
7	EFP-1	When studying, I work as hard as possible	☐1	☐2	☐3	☐4
8	SE-2	I'm confident I can understand the most complex material presented by the teacher	☐1	☐2	☐3	☐4
9	EL-1	When I study, I try to relate new material to things I have learned in other subjects	☐1	☐2	☐3	☐4
10	MEM-3	When I study, I memorise all new material so that I can recite it	☐1	☐2	☐3	☐4
11	CE-2	If I decide not to get any bad grades, I can really do it.	☐1	☐2	☐3	☐4
12	EFP-2	When studying, I keep working even if the material is difficult	☐1	☐2	☐3	☐4
13	CS-2	When I study, I force myself to check to see if I remember what I have learned.	☐1	☐2	☐3	☐4
14	IMOT-2	I study to ensure that my future will be financially secure.	☐1	☐2	☐3	☐4
15	MEM-4	When I study, I practice by saying the material to myself over and over	☐1	☐2	☐3	☐4

16	CE-3	If I decide not get any problems wrong, I can really do it.	☐1	☐2	☐3	☐4
17	EL-2	When I study,. I figure out how the information might be useful in the real world	☐1	☐2	☐3	☐4
18	SE-3	I'm confident I can do an excellent job on assignments and tests	☐1	☐2	☐3	☐4
19	CS-3	When I study, I try to figure out, as I read, which concepts I still haven't really understood.	☐1	☐2	☐3	☐4
20	EFP-3	When studying, I try to do my best to acquire the knowledge and skills taught	☐1	☐2	☐3	☐4
21	EL-3	When I study, I try to understand the material better by relating it to things I already know.	☐1	☐2	☐3	☐4
22	IMOT-3	I study to get a good job	☐1	☐2	☐3	☐4
23	CS-4	When I study, I make sure that I remember the most important things.	☐1	☐2	☐3	☐4

			almost never	some-times	often	almost always
24	CE-4	If I want to learn something well, I can	☐1	☐2	☐3	☐4
25	EL-4	When I study I figure out how the material fits in with what I have learned	☐1	☐2	☐3	☐4
26	SE-4	I'm certain I can master the skills being taught	☐1	☐2	☐3	☐4
27	CS-5	When I study, and I don’t' understand something, I look for additional information to clarify the point.	☐1	☐2	☐3	☐4
28	EFP-4	When studying, I put forth my best effort	☐1	☐2	☐3	☐4

			dis-agree	dis-agree some-what	agree some-what	agree
29	IMOTm-1	When I do math, I sometimes get totally absorbed	☐1	☐2	☐3	☐4
30	COOL-1	I like to work with other students	☐1	☐2	☐3	☐4
31	SCa-1	I learn things quickly in most school subjects	☐1	☐2	☐3	☐4
32	COML-1	I like to try to be better than other students.	☐1	☐2	☐3	☐4
33	SCv-1	I'm hopeless in English classes (R)	☐1	☐2	☐3	☐4
34	IMOTv-1	Because reading is fun, I wouldn't want to give it up	☐1	☐2	☐3	☐4

35	Sca-3	I'm good at most school subjects	☐1	☐2	☐3	☐4
36	COOL-2	I learn the most when I work with other students	☐1	☐2	☐3	☐4
37	SCv-2	I learn things quickly in English class	☐1	☐2	☐3	☐4
38	IMOTm-3	Because doing math is fun, I wouldn't want to give it up.	☐1	☐2	☐3	☐4
39	COML-2	Trying to be better than others makes me work well	☐1	☐2	☐3	☐4
40	SCm-1	I get good marks in mathematics	☐1	☐2	☐3	☐4
41	IMOTv-2	I read in my spare time	☐1	☐2	☐3	☐4
42	COOL-3	I do my best work when I work with other students	☐1	☐2	☐3	☐4
43	SCm-2	Mathematics is one of my best subjects	☐1	☐2	☐3	☐4
44	COML-3	I would like to be the best at something	☐1	☐2	☐3	☐4
45	IMOTv-3	When I read, I sometimes get totally absorbed	☐1	☐2	☐3	☐4
46	SCm-3	I have always done well in mathematics	☐1	☐2	☐3	☐4
47	COOL-4	I like to help other people do well in a group	☐1	☐2	☐3	☐4
48	SCa-2	I do well in tests in most school subjects	☐1	☐2	☐3	☐4
49	IMOTm-2	Math is important to me personally	☐1	☐2	☐3	☐4
50	COOL-5	(it is helpful to put together everyone's ideas when working on a project)	☐1	☐2	☐3	☐4
51	SCv-3	I get good marks in English	☐1	☐2	☐3	☐4
52	COML-4	I learn faster if I'm trying to do better than the others	☐1	☐2	☐3	☐4

CODES

MEM	Memorising
EL	Elaboration
CS	Control Strategies
IMOT	Instrumental Motivation
IMOTm	Instrumental Motivation Mathematics
IMOTv	Instrumental Motivation Verbal
EFP	Effort and Persistence in learning
COOL	Cooperative Learning
COML	Competitive Learning
CE	Control Expectation
SE	Self-Efficacy
SCv	Self-concept verbal
SCm	Self-concept math
Sca	Self-concept academic

References

Baumert, J., Fend, H., O'Neil, H.F., Peschar, J.L. (1998) Prepared for Life-Long Learning. Frame of Reference for the Measurement of Self-Regulated Learning as a Cross-Curricular Competency (CCC) in the PISA Project. Paris: OECD.

Byrne, B.M. (1996) Measuring Self-Concept Across the Life Span. Issues and Instrumentation.Washington D.C.: American Psychological Association.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. In: Structural Equation Modeling, vol 6, 1-55.

Ellis, J.L., & Wollenberg, A.L. van den (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. Psychometrika, 58, 417 - 429.

Grayson, D.A. (1988). Two-group classification in latent trait theory: Scores with monotone likelihood ratio. Psychometrika, 53, 383 - 392.

Hemker, B.T., & Sijtsma, K. (1993). A practical comparison between the weighted and the unweighted scalability coefficients of the Mokken model. Kwantitatieve Methoden, 14, no. 44, 59 - 73.

Junker, B.W. (1993). Conditional association, essential independence and monotone unidimensional item response models. The Annals of Statistics, 21, 1359 - 1378.

Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of 'scale analysis' and factor analysis. Psychological Bulletin, 45, 507 - 530.

Marsh, H. W., Craven, R. G., McInerney, D., & Debus, R. L. (1999). Evaluation of the Big-Two-Factor Theory of Motivation Orientation: Higher-order Factor Models and Age-related Changes. SELF Research Centre, University of Western Sydney, Macarthur.

Molenaar, W., (1997), Nonparametric models for polytomous responses. In: Linden, W.J. van der, Hambleton, R.K., (eds.), Handbook of Modern Item Response Theory. New York: Springer, 369-380.

Mokken, R.J., (1997), Nonparametric Models for Dichotomous Responses. In: Linden, W.J. van der, Hambleton, R.K., (eds.), Handbook of Modern Item Response Theory. New York: Springer, 351-368.

Niessen, M., Peschar, J.L. (ed.) (1982) Comparative Research on Education. Overview, Strategy and Applications in Eastern and Western Europe. Oxford: Pergamon Press /Budapest: Akademiai Kiado.

OECD (1996 ff.) Education at a Glance. Paris: OECD.

OECD (1997) Prepared for Life? How to Measure Cross-Curricular Competencies. (Bilingual). Paris: OECD.

Peschar, J.L., Molenaar, I.W. (1999) Plan of Analysis of Self-Regulated Learning as a Cross-Curricular Competency in PISA's Field Study. University of Groningen. Department of Sociology, Department of Statistics, Measurement Theory and Information Technology.

Robinson, J.P., Shaver, P.R., Wrightsman, L.S. (1991) Measures of Personality and Social Psychological Attitudes. San Diego: Academic Press.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425 - 435.

Rosenbaum, P.R. (1987). Comparing item characteristic curves. Psychometrika, 52, 217 - 233.

Stout, W.F., (1990). A New Item Response Theory Modeling Approach With Applications to Unidimensionality Assessment and Ability Estimation. Psychometrika, 55, 293-326.

TAG (1999) Technical Advisory Group Field Trial Design and Analysis Plan. Paper NPM (98) 9, PISA Consortium.

Verhelst, N.D., Glas, C.A.W. & Verstralen, H.H.F.M. (1995) One-Line Parameter Logistic Models (OPLM), Arnhem, Cito.

Wu, L.M., Adams, R.J. & Wilson M.R. (1998) ACER ConQuest: General Item Response Modeling Software. Camberwell, Australia: ACER Press.