THE IEA 10-YEAR TREND STUDY OF READING LITERACY: A MULTIVARIATE REANALYSIS

Jan-Eric Gustafsson, Gothenburg University, Sweden
Monica Rosén, Gothenburg University, Sweden

Abstract

In the paper the data from the 9 countries participating in the IEA 10-Year Trend Study of reading literacy was reanalyzed using multivariate methods. In the first step a three-factor latent-variable model was fitted to the scores from the 15 textblocks included in the reading test. One factor was a reading comprehension factor, another was interpreted as a speed factor, and the third factor reflected performance on documents texts. In the next step differences in latent variable means were investigated for the countries. This was done in two different ways: through multiple-group modelling and through computing means on estimated factor scores. The results obtained in these analyses were compared with one another, and with the estimated IRT scores on narrative, expository and documents test types. The most important finding was that the speed factor identified in the latent variable model affected the estimates of the changes in level of reading comprehension over time, as well as the estimated differences between countries. Differences also were noted in the results obtained for the speed factor when using estimated factor scores and results obtained using estimated latent variable means.

INTRODUCTION

In 1991 some 30 countries participated in a study of achievement in reading literacy organized by IEA (Elley, 1994). This study, which is referred to as the IEA Reading Literacy study (RL 1991), comprised one population of 9-10 year old students and another population of 14-15 year old students. In 2001 a subset of 9 of these countries (Greece, Hungary, Iceland, Italy, New Zealand, Singapore, Slovenia, Sweden and the United States) participated in a repeat study (RL 2001) in which the same instruments were administered to samples of the younger population of students. The students attented either grade 3 or grade 4, and in almost all cases the students from a particular country attended the same grade in the RL 1991 and RL

2001 studies. The comparison of results achieved in 1991 and 2001 is referred to as the 10 Year Trend study, and the design and findings have been reported by Martin, Mullis, Gonzalez, & Kennedy (2003).

The trend study showed that reading performance had improved in several countries, and most strongly so in Iceland, Greece, Slovenia and Italy. For Sweden results were poorer in 2001 than in 1991, and for the United States performance was unchanged. While Sweden was among the best-performing countries in 1991 that was not the case in 2001, when Italy, Iceland and the United States were the three best-performing countries.

Rosén and Gustafsson (2003) made a somewhat closer analysis of the Swedish data, using multivariate techniques. The reanalysis took advantage of the fact that the reading tasks included in the assessment represented texts from three different domains: narrative texts, expository texts and documents. A three-dimensional confirmatory factor analysis (CFA) model was fitted to the scores on the items associated with the 15 text passages included in the reading assessment (Gustafsson & Rosén, 2003). One factor was a general reading comprehension factor, with the highest relation to performance on narrative and expository text blocks. Another factor was related to performance on the last few blocks of each of the two separately timed booklets of reading tasks, and it was interpreted to reflect reading speed. The third factor was a documents factor, related to text blocks using noncontinous forms of presentation in graphs, tables, lists and so on.

The reanalysis showed that for Sweden there was a slight decrease in the level of reading comprehension. For reading speed a larger decrease was observed, while for the documents factor there was an increase in level of performance. Thus, while the overall pattern of findings agreed with those of the original analysis, the multivariate analysis provided a more complex and nuanced picture of the pattern of results. The results for the reading speed factor in particular, which was not included in the original analysis, contributed to this. In a similar reanalysis of the RL 1991 data, Gustafsson (1997) also showed that when reading speed was taken into account the rank ordering of countries according to reading literacy achievement was affected.

These results suggest that it would be interesting to reanalyze the data from the trend study in a similar way as has been done for Sweden. The purpose of the present paper is to apply such a multivariate analyses to the data from all 9 countries that participated in the trend study.

METHOD

Below the variables, subjects and methods of analysis are described.

The reading tasks

Elley (1994, p. 5) defined reading literacy as *"... the ability to understand and use those written language forms required by society and/or valued by the individual."* This definition thus makes a basic distinction between such forms of written language which citizens of a modern society encounter and must cope with (e. g., directions, maps, graphs, and government circulars) on the one hand, and leisure

reading (e. g., narrative prose) on the other hand. The definition also puts an emphasis on both "understanding" and "use", even though Elley (1994, p. 6) also pointed out that the constraints of the assessment situation imply that the focus is placed on "understanding."

On the basis of this definition of reading literacy, three different kinds of texts were selected for inclusion in the RL instruments (Elley, 1994, p. 6):

- *Narrative Prose.* These are continous texts which aim to tell a story. The texts typically follow a linear time sequence, and are usually intended to entertain or to involve the reader emotionally. The texts ranged in length from short fables to longer stories.
- *Expository Prose.* This category comprises continous texts which aim to convey factual information or opinion to the reader. Descriptions of animals is a frequently used content of these texts.
- *Documents.* These are structured presentations of information, in the form of graphs, charts, maps, lists, or sets of instructions. The reader can process the information in a non-linear fashion without reading the whole text, and typically the number of words is limited.

Four narrative, five expository, and six documents text passages were developed for inclusion in the instrument. These texts, along with the accompanying items, were organized into two booklets (Booklet A and Booklet B), each booklet including a mixture of the three categories of texts.

According to the instrument design each participating student was expected to complete both Booklet A and Booklet B, and normally these were administered on the same day. In Sweden, however, the two booklets were administered at two occasions about one week apart. The allowed testing time for the text passages included in Booklet A was 35 minutes, while it was 40 minutes for Booklet B. The overall testing time thus was 75 minutes for the RL study.

The tests included altogether 66 items: 22 narrative, 21 expository, and 23 documents items. Most of the items were of the multiple-choice type. The items were designed to measure different kinds of processes (Elley, 1994, pp. 11-12): 8 items required a verbatim response; 20 items required the student to paraphrase the answer in different wording from the question; 15 items required students to go beyond the information given and make an inference; 11 items required students to locate a fact or figure; and 12 items asked them both to locate and process information.

Subjects

Within the countries the same basic two-stage stratified cluster sampling design was employed, even though there were differences in the details of the implementation. In the first step, schools were sampled according to the PPS-principle (probability proportional to size), in most cases with stratification of schools into explicit strata. One classroom in each school was then sampled for inclusion in the study. As is described in the PIRLS Technical Report (Martin, Mullis & Kennedy, 2003), sample

weights have been computed to allow inferences to the population, given the characteristics of the particular sampling design used. Appendix A of the Mullis et al. (2003) international report provides detailed information about school and student participation rates. For all countries, except Greece and the United States, school participation rates exceeded 93 %. Student participation rates exceeded 93 % for all countries except three: Greece, Iceland and the United States.

Some students did not take either Booklet A or Booklet B, causing missing data for some of the items. While the multivariate techniques employed here in certain instances can deal with partially missing data, the multiple-group modelling methods employed here cannot do that. It was therefore necessary to use listwise-deletion of missing data, thus including only students with data on both booklets. This caused some reduction of the sample size, and it implies that results presented here are not identical with those presented in the international report (Martin et al., 2003). Table 1 presents the number of students included in the analyses.

Table 1: Number of Students Included in the Analyses

Country RL 1991 RL 2001 Total			
Greece	3500	1108	4608
Hungary	2914	4702	7616
Iceland	3932	1783	5715
Italy	2221	1589	3810
New Zealand	2960	1179	4139
Singapore	7326	3600	10926
Slovenia	3297	1502	4799
Sweden	4245	5075	9320
United States	6365	1822	8187
Total	36760	22360	59120

As can be seen in the table the samples for RL 2001 tend to be smaller than those for RL 1991. The main reason for this is that the suggested design of the repeat study only encompassed a single class in each of 80 schools. However, a few countries sampled a larger number of schools and/or included all classes in the schools. Altogether the two studies included about 59 000 students, of which around 37 000 were in RL 1991 and 22 000 were in RL 2001.

It should be pointed out that the number of individual students only gives approximate information about the amount of information available in the data, given that cluster sampling was used. In such a design the actual amount of information is a function not only of the number of individual observations, but also of the cluster size and the amount of intra-cluster similarity, as expressed by the intraclass correlation. In several of the school-systems under study here the school-level intraclass correlations are substantial, which implies that the actual sample size

is smaller than is indicated by the nominal sample size. The ways in which this problem were dealt with are described below.

Variables

In the modelling of the data variables aggregated to the passage level have been used. There were 15 such passage scores. The variables were labeled in such a way that the first character identifies the text type (D = documents, E = expository, and N = narrative), the second character the booklet in which the block appeared (1 or 2), and the following three characters are an abbreviation of the label of the block. All the RL items were scored 0 or 1, so the maximum score is also the number of items in the passage.

A few items had to be excluded because they were missing for one or more of the countries at one of the two occasions (ADBUSES3 and ANBIRD04), or because there were problems identifying a single correct answer (ADTEMPR2 and ADMARIA02). Two items requiring open-ended answers also were excluded because they had not been scored in all countries (ANGRAPA07 and AEWALRU7). Table 2 presents descriptive statistics for the 15 passage variables.

Table 2: Descriptive Statistics for the Variables

Passage	*N*	*Minimum*	*Maximum*	*Mean*	*Std Dev*
E1CRD	59122	0	2	1.87	0.38
N1BRD	59122	0	4	2.74	1.11
D1ISL	59122	0	4	2.99	1.02
D1MRA	59122	0	2	1.71	0.54
N1DGS	59122	0	6	3.90	1.77
E1WLR	59122	0	6	4.54	1.71
E2SND	59122	0	3	2.29	0.91
N2SHK	59122	0	5	3.75	1.36
D2BTT	59122	0	4	3.26	1.01
D2BUS	59122	0	3	1.99	1.03
D2CNT	59122	0	3	2.66	0.74
D2TMP	59122	0	4	2.32	1.18
E2MRM	59122	0	4	1.92	1.27
N2GRP	59122	0	6	4.02	2.04
E2TRE	59122	0	6	2.86	1.80

It may be observed that for most of the variables the means tend to be fairly high in relation to the maximum possible score, indicating a ceiling effect.

Methods of Analysis

CFA is a powerful method for decomposing a larger set of observed variables into a more limited set of latent variables (see, e. g., Kline, 1998; Loehlin, 1998). However, while the basic principles of estimating and and testing a CFA model for a single group of randomly sampled cases are well known and relatively simple, the present data offers some special challenges.

One of these challenges is that the data include two samples from each of nine countries, or 18 different groups. The large number of groups may be dealt with in different ways. One possibility is to represent group membership as a vector of 17 dummy-variables, each dummy variable representing either the RL 1991 study or the RL 2001 study for a particular country. In this case a model is fitted to a single covariance matrix. This approach allows estimation of differences in level of performance on the latent variables for the 17 different groups as compared to a reference group. However, this approach implies an assumption that all the parameters of the measurement model (i.e., factor loadings, error variances in manifest variables, and variances in latent variables) are identical for the 18 groups.

Another possibility is to approach the problem with multiple-group modelling techniques. In this approach one model is fitted to each of the covariance matrices for the 18 groups. However, constraints of equality over groups may be imposed on some or all of the parameters of the measurement model. In one extreme every parameter is constrained to be equal to the corresponding parameter in every other group. When differences in latent variable means are investigated within such a completely constrained model, the same estimates of group differences in level of performance are obtained as when the one-group dummy-variable approach is used. However, if the constraints imposed do not agree with the group differences that are present in the data the model fit will be poor. Thus, while group differences in the measurement properties of the reading tasks cannot be detected in the dummy-variable approach, such differences can appear in the multiple-group approach. If a model which imposes constraints of equality of parameters over groups does not fit the data these constraints may be relaxed. In this way differences between groups with respect to the measurement properties of the reading tasks may be investigated.

While both the dummy-variable approach and the multiple-group modelling approach have advantages in terms of parameter estimation and testing of hypotheses, these approaches can become overwhelmingly complex. This is particularly true of the multiple-group approach, and especially so if it is extended into full-fledged structural equation models (SEM) by including further variables as explanatory or control variables. These complexities may be avoided in yet another approach, which is a two-step procedure. In the first step factor scores are computed for each individual and these factors scores are in the next step used in analyses, using either SEM methodology or other techniques such as HLM (Bryk & Raudenbush, 2002). The factor score approach thus has advantages from practical points of view. However, the individual estimates of factor scores are not error-free, which is the case with latent variables within CFA models.

Other challenges in analyzing the current data with multivariate techniques are

caused by the sampling design. The stratification makes it necessary to take the case weights into account and all SEM programs cannot do that. However, one of the programs that does take weights into account is the Mplus program (Muthén & Muthén, 2002) and it has been used for the analyses in the present paper. As has already been mentioned, the fact that cluster sampling was also used creates problem. The two- (or three-) level nature of data could, in principle, warrant using multi-level analytical techniques. While certain SEM programs (e. g., Mplus) can fit latent-variable models for two-level data it would be a complex undertaking indeed to perform such analyses for 18 groups. However, within the Mplus 2 program there is a "complex sample" option which for a one-level model adjusts the estimates of standard errors and goodness-of-fit statistics according to the cluster design. At present this option is not available for multiple-group models, which limits its applicability in the present study. Nevertheless, in some of the analyses to be reported here this option has been used to correct for the effects of cluster sampling.

The modelling was carried out with Mplus 2 (Muthén & Muthén, 2001) under the STREAMS 2.5 modelling environment (Gustafsson & Stahl, 2000).

RESULTS

The analysis of data proceeded in three steps. In the first step alternative CFA models were compared. In the next step multiple-group models were fitted, and differences in model parameters over groups were investigated. In the final step differences in latent variable means were investigated for the countries. This was done in two different ways: through multiple-group modelling and through computing means on estimated factor scores.

Alternative measurement models

Gustafsson & Rosén (2003) fitted a model with three factors to the 15 passage variables for the Swedish RL 2001 data (see also Gustafsson, 1997; Yang, 1998). One of the variables was a general reading comprehension factor with relations to all 15 passage scores. Another factor was a (residual) documents factor with relations to the six Documents passage scores. The third factor was an "end of test" factor with relations to the last few passages of the two separately timed booklets. All three factors were orthogonal to one another, and since the model includes a general factor this is a hierarchical model of the so-called nested-factor type (see Gustafsson, 2001). Attempts to fit this model to data from other countries were not entirely successful, however, since the model failed to converge for some of the sets of data. This suggests that alternative model structures need to be tried. On the basis of previous modelling experiences, a set of increasingly complex models were therefore postulated, as follows:

1. A *one-factor* model with a single factor which was related to all 15 passage scores.
2. A *two-factor* model with two correlated factors, one of which related to the nine narrative and expository passage scores, and the other to the six documents passage scores.
3. A *three-factor* model which included the same two factors as the two-factor

model, along with an "end of test" factor related to the two last passages of Booklet A and the four last passages of Booklet B. In this model the "end of test" factor was uncorrelated with the other two factors.

4. A *four-factor* model which included three correlated factors, one related to the narrative passage scores, another related to the expository passage scores, and the third related to the documents passages. The fourth factor was the same factor as the "end-of-test" factor in the three-factor model, and it was defined to be orthogonal to the other three factors.

These four models were fitted to the RL 1991 and RL 2001 data for the 9 countries as multiple-group models of two different kinds. In one kind of model every parameter of the measurement model, including latent variable means, was constrained to be equal over the 18 groups, and in the other kind of model no constraints of equality over groups were imposed. Goodness-of-fit statistics for the eight models are presented in Table 3.

Table 3: Goodness-of-Fit Statistics for Alternative Multiple-Group Measurement Models

Model	*Chi-square*	*Df*	*CFI*	*RMSEA Δ*	*Chi-square*	*Δ Df*
1 factor, constraints	92211.3	2385	0.690	0.107		
2 factors, constraints	86584.5	2384	0.709	0.104	5626.8	1
3 factors, constraints	77993.3	2377	0.739	0.098	8591.2	7
4 factors, constraints	77303.1	2375	0.741	0.098	690.2	2
1 factor, no constraints	24997.0	1620	0.919	0.066		
2 factors, no constraints	18301.9	1602	0.942	0.056	6695.1	18
3 factors, no constraints	8273.8	1476	0.977	0.037	10028.1	126
4 factors, no constraints	7662.0	1440	0.979	0.036	611.9	36

The one-factor model had a poor fit, both with and without constraints of equality over groups. Introducing the distinction between the documents passages and the passages involving comprehension of continous text in the two-factor model caused a considerable improvement of fit. However, the RMSEA index of .056 in the non-constrained model indicates that there is room for improvement of fit. Introducing the EoT factor in the three-factor model caused a substantial drop in the χ^2 statistic in both the constrained and the non-constrained models. For the non-constrained model the RMSEA index was .037 and the CFI index was .977, which indices both indicate excellent fit between model and data. Introducing the distinction in the four-factor model between the narrative and expository domains of text caused some reduction in the χ^2 statistic. However, according to both the RMSEA and CFI indices the improvement was so marginal that it does not seem warranted to make the distinction in the model between the two types of continous text.

Model differences between groups

These model comparisons suggested that the three-factor model was the one to be preferred, so this model was investigated more closely. Table 4 presents goodness-of-fit statistics for the three-factor model for each of the 18 groups of cases. The model was estimated in two different ways: with the complex sample option in Mplus, to take the clustering of students into schools into account, and as an ordinary CFA model assuming independently sampled observations.

Table 4: Goodness-of-Fit Statistics for the Three-Factor Model for the Different Groups

Country	*Study*	*N*	*Complex sample*			*Non-complex sample*		
			χ^2	*Df*	*RMSEA*	χ^2	*Df*	*RMSEA*
Greece	1991	3498	358.5	82	0.031	794.6	82	0.050
Greece	2001	1108	167.9	82	0.031	210.8	82	0.038
Hungary	1991	2911	249.6	82	0.027	343.1	82	0.033
Hungary	2001	4701	355.1	82	0.027	563.2	82	0.035
Iceland	1991	3932				447.7	82	0.040
Iceland	2001	1784				259.5	82	0.035
Italy	1991	2221	152.7	82	0.020	283.1	82	0.033
Italy	2001	1589	138.2	82	0.021	193.6	82	0.029
New Zealand	1991	2960	330.3	82	0.032	523.5	82	0.043
New Zealand	2001	1181	192.5	82	0.034	281.4	82	0.045
Singapore	1991	7326	614.5	82	0.030	815.2	82	0.035
Singapore	2001	3599	452.3	82	0.035	619.5	82	0.043
Slovenia	1991	3297	303.3	82	0.029	358.2	82	0.032
Slovenia	2001	1502	193.3	82	0.030	270.5	82	0.039
Sweden	1991	4239	471.9	82	0.033	679.9	82	0.041
Sweden	2001	5101	513.3	82	0.032	854.2	82	0.043
United States	1991	6367	360.9	82	0.023	459.0	82	0.027
United States	2001	1821	208.6	82	0.029	315.0	82	0.039

For reasons which are not clear, Mplus failed to compute a correct solution under the complex sample option for the two Icelandic samples. However, for all other groups a solution could be obtained under both the complex and non-complex sampling assumptions. As may be expected the χ^2 statistic was in all cases lower when the cluster sampling was taken into account. The sum of the χ^2 statistics of all the groups is the χ^2 statistic obtained when estimating a multiple-group model

without any equality constraints over groups. This makes it possible to determine the amount of overestimation caused by the cluster sampling in multiple-group models, even though Mplus does not allow estimation of multiple-group models under the complex sample option. For all groups except for the two Icelandic groups the sum of χ^2 statistics under the non-complex option was 7564.7 with 1312 df, while under the complex option the corresponding sum was 5062.9. Thus, for the present data the cluster sampling caused the χ^2 statistic to be overestimated by about 50 % when independently sampled observations were assumed.

From Table 4 it may be observed that the χ^2 statistic was highly significant in all cases. However, large sample sizes were employed, and as may be seen in the table there is quite a close association between the sample size and the size of the χ^2 statistic. The RMSEA statistic corrects for the differences in sample size and the RMSEA estimates are quite homogeneous and low, with values around .03. These results show that the fit of the three-factor model was excellent for all groups.

The three-factor model with constraints of equality over groups had a very poor fit, however (see Table 3). This shows that there were differences in the parameter estimates of the different groups. There are, however, several categories of parameters for which there may be differences between the groups. A series of models were therefore compared, in which the constraints of equality over groups were successively removed for groups of parameters. Table 5 summarizes the results of the model comparisons.

Table 5. Tests of Fit of Models with Different Constraints of Equality over Groups

Models	χ^2	*df*	*RMSEA*	$\Delta\chi^2$	Δdf	$\Delta\chi^2/\Delta df$
1. Equality constraints over groups	77933.1	2377	0.098			
2. No constraints on latent variable means	71514.8	2326	0.095	6418.3	51	125.8
3. No constraints on manifest variable means	41813.4	2122	0.075	29701.4	204	145.6
4. No constraints on error variances	22801.2	1867	0.058	19012.2	255	74.6
5. No constraints on factor variances and covariances	16444.2	1799	0.050	6357	68	93.5
6. No constraints on factor loadings for ReadC	13555.7	1663	0.047	2888.5	136	21.2
7. No constraints on factor loadings for Doc	10345.9	1578	0.041	3209.8	85	37.8
8. No constraints on factor loadings for EoT	8273.9	1476	0.037	2072	102	20.3

The starting point was the three-factor model with constraints of equality on all parameters over all groups (Model 1). In the first step the constraints of equality on latent variable means were relaxed (Model 2), which caused a considerable reduction

in the size of the χ^2 statistic, the decrease amounting to 125.8 χ^2 units/df. While it must be remembered that this statistic is overestimated because the effects of the cluster sampling have not been taken into account, the improvement was strong enough to show that there are differences in level of performance on the latent variables over the groups. In Model 3 the constraints of equality on the manifest variable means were relaxed. Thus in this model differences in the means of the observed variables over and above those due to the latent variable mean differences were allowed. This relaxation of constraints too caused an improvement of fit, which amounted to 145.6 χ^2 units/df. This result thus shows that the pattern of performance differences between the groups on the different passages was only partially accounted for by the differences in latent variable means.

In Model 4 the constraints of equality were relaxed for the error variances of the manifest variables. This caused fit to improve by about 75 χ^2 units/df (see Table 5), which showed that there were differences in the error variances for the groups. Model 5 took the further step of relaxing constraints of equality on the variances of the three latent variables and the covariance between ReadC and Doc, which caused a considerable improvement of fit (93.5 χ^2 units/df). After these constraints were relaxed the overall fit of the model may be regarded as acceptable, with an RMSEA estimate of 0.050. In Models 6, 7 and 8 the equality constraints on the factor loadings (i. e., relations between latent and manifest variables) were relaxed for the ReadC, Doc and EoT latent variables, respectively. All three models brought about improvements of fit, but measured in terms of χ^2 units/df the relaxations of the equality constraints on the factor loadings caused less improvements of model fit than the other relaxations of equality constraints. After the constraints of equality on the factor loadings of the Eot factor were removed no constraints remained, so Model 8 is identical with the unconstrained three-factor model (see Table 3).

Summarizing the findings from these model comparisons it may be concluded that there were differences between the models for the groups in many different respects, such as means on latent and manifest variables, variances of errors in manifest variables, variances of latent variables and factor loadings. A complete account of the differences between the 18 groups would thus require a very detailed description. However, such a detailed description could easily become overwhelmingly complex, and the major pattern of differences between groups might be lost in the details. This may be a reason in a first step to base the description of the group differences on a constrained model, and then possibly allow the complexities of a less constrained model to appear. Given the focus on group differences in level of performance, it would seem that a closer analysis of group differences at the passage score level would be an interesting complement to the analysis of differences in latent variable means.

While the results presented here showed that there were some differences between groups with respect to the factor loadings, these differences seemed to be smaller than for the other categories of parameters. Given that the χ^2 statistics in the present study were inflated by the large sample size and the cluster sampling design there is reason to be cautious in concluding that there were group differences in factor

loadings. According to the RMSEA index, the overall fit of the model with equality constraints on factor loadings was quite acceptable, which suggests that the differences were of relatively small magnitude.

Group differences in factor means

In order to investigate differences between the groups in the latent variable means, two different approaches were employed. In one approach the estimated latent variable means for the groups in Model 2 were compared, and in the other approach factor scores were first estimated, which were then analyzed with respect to group differences. The factor scores were estimated from the three-factor model applied on the complete data, pooled into one data set. Using the senate weights (see Martin, Mullis, & Kennedy, 2003), which weigh the participating countries equally, the RL 2001 factor scores were scaled to have a mean of 500 and a standard deviation of 100.

It is interesting to compare the results obtained with the CFA approach with those obtained with the IRT approach used in the original reporting of results (Martin et al., 2003). Table 6 presents the results obtained with the overall Reading Literacy IRT score and the Documents IRT score.

Table 6: Mean Estimated IRT scores for RL 1991 and RL 2001

Country	*Reading Literacy IRT Score*			*Documents IRT Score*		
	RL 1991	*RL 2001*	*Change*	*RL 1991*	*RL 2001*	*Change*
Greece	466	507	40	442	492	50
Hungary	461	476	15	467	486	18
Iceland	486	513	27	481	506	25
Italy	501	513	11	482	499	17
New Zealand	499	503	4	492	508	16
Singapore	481	488	7	465	484	19
Slovenia	458	494	36	454	503	49
Sweden	514	501	-13	504	508	4
United States	522	511	-11	526	519	-7

The results agreed quite well with those presented in the international report (Martin et al., 2003), even though there also were some slight differences which are due to the fact that the results reported here only include those students who have complete data. The largest difference was observed for the Swedish RL 2001 data, where it amounted to 3 score points. However, the main purpose here is to compare the different methods of analysis, so for the moment we will disregard these differences. Table 7 presents results obtained when the analysis was based on the estimated factor scores.

Table 7: Estimates of Factor Means Based on Factor Scores

	ReadC Factor Score			*Doc Factor Score*			*EoT Factor Score*		
Country	*RL 1991*	*RL 2001*	*Change*	*RL 1991*	*RL 2001*	*Change*	*RL 1991*	*RL 2001*	*Change*
Greece	467	511	44	454	502	47	518	508	-10
Hungary	461	481	20	467	485	19	488	474	-14
Iceland	485	512	28	481	510	29	493	495	2
Italy	495	508	13	489	506	17	523	521	-3
New Zealand	496	499	2	494	500	5	512	514	3
Singapore	476	480	4	473	481	8	530	529	-1
Slovenia	462	499	37	459	503	44	472	484	11
Sweden	504	494	-9	500	496	-4	508	483	-24
United States	523	516	-7	525	518	-7	501	491	-9

It may first of all be noted that while for most countries there was an increase in level of performance on the ReadC and Doc factors between 1991 and 2001, for almost all countries there was a decrease in level of performance on the EoT factor. The results for the ReadC factor score are most directly comparable with the overall Reading Literacy IRT score, and the estimates agreed quite well. For some countries there was, however, a tendency for the factor scores to be higher than the IRT scores. This held true for Greece, Hungary, Sweden and the United States. For these four countries there was also a large negative change in the EoT factor between 1991 and 2001. The Reading Literacy IRT score reflects the combined effect of increases in reading comprehension and decreases in reading speed, while these changes are separated in the estimated factor scores. The differential pattern of change for reading comprehension and reading speed for these four countries may thus account for the somewhat different pattern of results for the overall IRT score and the ReadC factor score.

The Documents IRT score and the Doc factor score also showed good agreement for most countries. However, for New Zealand, Singapore and Sweden the factor score estimate was lower by 8 to 11 score points. This may be due to the fact that in the computation of the factor score the single passages are weighted somewhat differently than when the IRT score is computed. Another reason may be that three documents items were excluded in the analyses reported here.

We will next turn to the results obtained when the estimates are based directly on the multiple-group CFA model. In these analyses the latent variable means for Greece RL 1991 have been taken to be zero, while the latent variable means for the all other groups have been estimated as free parameters. The differences have been standardized with the within-group standard deviations, so the results are expressed on a d-scale. Table 8 presents the estimated latent variable means.

Table 8: Estimated Latent Variable Means for RL 1991 and RL 2001

	ReadC			*Doc*			*EoT*		
Country	*RL 1991*	*RL 2001*	*Change*	*RL 1991*	*RL 2001*	*Change*	*RL 1991*	*RL 2001*	*Change*
Greece	0	.50	0.50	0	.54	0.54	0	-.41	-0.41
Hungary	-.22	.08	0.30	.33	.53	0.20	-.05	-.53	-0.48
Iceland	.23	.53	0.30	.40	.71	0.31	-.59	-.71	-0.12
Italy	.28	.41	0.13	.43	.66	0.23	-.11	-.23	-0.12
New Zealand	.30	.34	0.04	.55	.65	0.10	-.28	-.41	-0.13
Singapore	-.05	.03	0.08	.30	.43	0.13	.43	.21	-0.22
Slovenia	.03	.39	0.36	.18	.72	0.54	-.87	-.89	-0.02
Sweden	.40	.32	-0.08	.59	.62	0.03	-0.4	-.74	-0.34
United States	.62	.61	-0.01	.93	.87	-0.06	-.74	-1.07	-0.33

Since the scales of the measures are different, no direct comparisons between the estimated factor scores may be made. However, since the estimate of the change has a standard deviation of 1.00 and the factor score has a standard deviation of 100 we can make a rough translation between the scales by multiplying the estimated change of the latent variable mean by 100.

For the ReadC factor the overall pattern of results agrees quite well with those obtained from the estimated factor scores. For some countries (Greece and Hungary in particular) there was, however, a tendency for the latent variable mean to have a higher estimate than the factor score. The largest differences were, however, observed for the EoT factor where the latent variable mean changes tended to show a stronger decrease than did the factor scores. This was particularly marked for Greece, Hungary, Singapore and the United States. One reason why we would expect stronger effects to be seen in the latent variable means than the estimated factor scores is that the variance of the latter is inflated by errors of measurement, which is not the case for the latent variable estimates. According to Mplus, the determinacy, which is a reliability estimate for factor scores, was .92 for ReadC, .91 for Doc, and .68 for EoT. The reliability thus is lower for EoT than for the other two factors. The reliability estimate of .68 for EoT implies that 32 % of the variance is error variance, which in turn implies that the change in means of factor scores is underestimated, in a similar way as a coefficient of correlation is attenuated by errors of measurement in the observed variables.

DISCUSSION AND CONCLUSIONS

According to the CFA modelling, three latent variables need to be differentiated to account for the covariances among the 15 passage scores. However, these three factors only partially overlap with the distinction between the three domains from which the reading tasks were sampled. Thus, expository and narrative tasks both loaded on the same factor, which may be interpreted as a reading comprehension factor. The tasks from the documents domain did however load on a separate

documents factor. The third factor in the model may be interpreted to reflect individual differences in reading speed, and possibly also motivation, and this factor does not have any counterpart in the previous reporting.

Even though the present analysis did not provide much support for the distinction between narrative and expository types of text there may be limits to the generalizability of this finding. Thus, for the passages included in the current study there does not seem to be a clearcut distinction between the two categories of text. However, for texts employed when older students are participating the distinction may be easier to make.

It should also be pointed out that even though the distinction between the two categories of text was not supported in the present analyses, there may be differences between school systems in the relative amount of emphasis put on the two categories of text. The effects of such differences in instructional emphasis are, however, not easily detected in the kind of analyses reported here, because the treatment effects are weak compared to the amount of individual differences within countries (see Gustafsson, 1997).

While the three-factor model fitted excellently to the data from all the groups, the multiple-group model also showed that there were differences in the size of the parameter estimates over countries and occasions. This seemed in particular to be the case for the intercepts of the manifest variables, the error variances in the manifest variables, and the factor variances and covariances. The differences with respect to the intercepts of the manifest variables show that in addition to the performance differences in latent variable means, different groups had advantages and disadvantages which were specific to particular texts. Within the framework of the current study it has not been possible to investigate these text-specific differences, but this is an urgent task for future research.

The results from the analyses based on estimated factor scores showed that for some countries there were larger improvements in reading comprehension between 1991 and 2001 than was indicated by the changes in the IRT score. The main reason for this seems to be that for these countries there was also a decrease in reading speed, and since the IRT score confounds reading comprehension and reading speed this causes a lower estimate. Another interesting finding was that for almost all the countries there was a decrease in reading speed between 1991 and 2001. The changes in the reading speed factor were brought out even more clearly when changes in the latent variable means were investigated within the multiple-group CFA model. One possible explanation for this is that the estimated factor scores are influenced by errors of measurement, which is not the case for the estimates of the latent variable means.

In summary, it seems that even though the overall pattern of results from the multivariate reanalysis agreed quite well with the originally reported results, the fact that the reading speed factor could be identified added further insight into the trends in reading literacy achievement between 1991 and 2001 for the nine participating countries.

References

Gustafsson, J-E. (1997). Measurement characteristics of the IEA reading literacy scales for 9- and 10-year-olds at country and individual levels. *Journal of Educational Measurement*, *34* (3), 233-251.

Gustafsson, J-E., & Rosén., M. (2003). The dimensional structure of reading assessment tasks in the IEA Reading Literacy Study 1991 and the Progress in International Reading Literacy Study 2001. Paper presented at the EARLI 10th Biennial Conference, Padova, Italy, August 26-30, 2003.

Gustafsson, J-E, & Stahl, P-A. (2000). *STREAMS User' s Guide, Version 2.5 for Windows 95/98/NT*. Mölndal, Sweden: MultivariateWare.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guildford Press.

Loehlin, J. C. (1998). *Latent variable models. An introduction to factor, path, & structural analysis (Third ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J. & Kennedy, A. M. (2003). Trends in Children's Reading Literacy Achievement 1991-2001: IEA Repeat in Nine Countries of the 1991 Reading Literacy Study. Chestnut Hill, Massachusetts: ISC Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.) (2003). PIRLS 2001 Technical Report. Chestnut Hill, Massachusetts: ISC Boston College.

Muthén, L. K. & Muthén, B. O. (2001). Mplus User's Guide. Los Angeles, CA: Muthén & Muthén.

Rosén., M., & Gustafsson, J-E. (2003). Changes in levels of reading literacy between 1991 and 2001; A trend study of 9-year olds in Sweden. Paper presented at the EARLI 10th Biennial Conference, Padova, Italy, August 26-30, 2003.