Issues in the measurement of change in reading achievement over time

Petra Lietz
The Flinders University of South Australia

The International Association for the Evaluation of Educational Achievement (IEA) conducted two major studies in which reading achievement was assessed, the Reading Comprehension Study in 1970/71 and the Reading Literacy Study in 1990/91. Eight countries, namely Belgium French, Finland, Hungary, Italy, the Netherlands, New Zealand, Sweden, and the United States of America participated in both surveys at the 14-year-old level. While the instruments that were employed differed considerably between the two studies, information was gathered on a number of common items in the reading tests as well as common questions in the student-, teacher- and school-questionnaires.
Thus, the data obtained in the two studies provide an unique opportunity to make comparisons over time with respect to the level of reading achievement of the 14-year-old students in the eight countries and the extent to which changes in the effects of background factors could have contributed to any changes in reading achievement that might have occurred over the 20-year-period.
However, certain issues need to be addressed before an attempt is made to equate the level of reading achievement over time. These issues include questions of the equivalence of the reading tests over time and across cultures and languages. Therefore, the analyses reported in this paper seek to examine:
(1)the factor structures underlying the tests employed in the Reading Comprehension and the Reading Literacy Studies. In particular, it is of interest whether or not the assumption holds of a common factor or trait, that underlies the processes involved in reading comprehension, as argued by Thorndike (1973a). If this were not the case, it would not be appropriate to calculate a total score which is necessary to equate student performance over time. Confirmatory factor analysis (CFA) is used to examine this issue.
(2)whether or not similar results are obtained across countries from an examination of the different factor structures. This is of importance in order to examine whether or not the reading comprehension of different languages in which the tests were administered would follow similar structures. If this were not the case, it would not be appropriate to make comparisons across countries. With respect to this issue, Hambleton and Kanjee (1994) have pointed out that

> such analyses [i.e. the study of factor structures] are central in assessing the equivalence of instruments across cultural/language groups. (Hambleton and Kanjee 1994, p. 6333)

(3)whether or not similar item characteristic may be obtained when the items from the two reading tests are scaled using the one-parameter

Rasch model, which assumes a common underlying trait. In this way, items may be identified which fit the scale poorly. A closer examination of the misfitting items, in turn, might reveal whether or not the poor fit could stem from problems that are associated with translation or the culturally specific context of an item. Again, this method has been noted by Hambleton and Kanjee (1994) to provide information on the equivalence of translated tests.

Factor structures underlying the tests employed in the Reading Comprehension and the Reading Literacy Study

Many researchers have attempted to examine the processes involved in

reading comprehension and whether or not distinct subskills might be identified and ordered hierarchically. While some analyses have generated results to support the existence of separate skills (Davis, 1968; O'Neill, 1978; Spearritt, 1972), which may then be ordered hierarchically (Clark, 1973; Ludlow and Hillocks, 1985), other results did not provide strong empirical evidence for the existence of multiple dimensions in reading comprehension (Reutzel and Hollingsworth, 1991; Thorndike, 1973a; Zwick, 1987). Indeed, while in the rationale for the objectives to be assessed in the reading tests of the National Assessment of Educational Progress in 1970/71 a hierarchical order of reading skills was assumed, the information on the objectives for the 1983/84 tests explicitly stated that no such hierarchy could be anticipated (NAEP 1971, 1984).

Other researchers (Bartlett, 1985; Elley, 1992, 1994; Guthrie, 1987) have suggested a distinction between different types of reading materials such as, for example, narrative and expository materials since different processes are involved in the reading comprehension of these materials. Thus, Guthrie (1987) argued that while all reading is likely to involve some form of decoding, the reading of a school timetable would be more concerned with locating information, while a novel would be likely to require literal comprehension and higher order processes.

These different assumptions were reflected in the design of the Reading Comprehension and Reading Literacy Studies.

In the Reading Comprehension Study, an attempt was made to include items that assessed student performance on the following reading skills which had originally been suggested by Davis (1944):

1.an ability to follow the organization of a passage and to identify antecedents and references in it;
2.an ability to answer questions that are specifically answered in the passage;
3.an ability to draw inferences from a passage about its contents; and
4.an ability to determine the writer's purpose or point of view.
(Thorndike, 1973b p. 56)

Items were assigned to subskills by expert judgement. It has to be noted, however, that not all 52 reading test items were categorized in this way with only 42 being assigned to a subskill. The remaining ten items were not assigned to any subskill.
In the analyses of subscores reported by Thorndike (1973b), the correlations between the four reading comprehension subscores were extremely high, indicating a high overlap of the processes involved when performing certain reading tasks. As a result, Thorndike (1973b) questioned the usefulness of the subscores calculated for the different skills and concluded:

… any attempt to distinguish between them [i.e. the subscores] will result in very fragile results. (Thorndike, 1973b p. 60)

In the Reading Literacy Study, great care was taken to design a test measuring three different domains of reading, namely Narrative, Expository and Documents. These domains were defined as:

1.Narrative prose:Continuous text in which the writer's aim is to tell a story - whether fact or fiction. They normally follow a linear time sequence and are usually intended to entertain or involve the reader emotionally. […]
2.Expository prose:Continuous text designed to describe, explain, or otherwise convey factual information or opinion to the reader. […]
3.Documents:Structured information displays presented in the form of charts, tables, maps, graphs, lists or sets of instructions. (Elley et al., 1992 p.4)

Elley et al. (1992) also stated:

From the outset, it was agreed that test scores would be reported separately for each of the three domains - Narrative, Expository and Documents. (Elley et al., 1992 p.4)

Accordingly, most of the reporting of bivariate relationships between student reading achievement and certain student, teacher and school variables was undertaken by providing separate figures for each domain. However, in the major summary report of the Reading Literacy Study (Elley et al., 1994 p. 12) it is noted, without further comment or discussion of the issue, that

… measures of student abilities in reading literacy were estimated and reported […] for each domain separately […] as well as for the total item scale.

Subsequent item analyses were undertaken using an international pooled dataset including all items from the different domains except those

which, according to the authors, had poor psychometric properties on the international scale. In addition, some of the results were reported for the total item scale, suggesting a certain ambiguity as to the test scores suitable for reporting results of the Reading Literacy Study, which would appear not to have been resolved.

In summary, there are conflicting views regarding the nature of the processes involved in reading comprehension. The designers of the Reading Comprehension Study set out to verify the existence of subskills that made different demands on the students' thought processes that were involved in answering questions to reading comprehension items. Their analyses, however, suggested that the processes overlapped to such a degree that a distinction between them would be meaningless.
In the design of the Reading Literacy tests, subscores were calculated for different types of reading materials on which the test items were based. While most of the Reading Literacy Study results reported were presented separately for the different subscores (Elley, 1992, 1994; Lundberg and Lynnakylä, 1993; Postlethwaite and Ross, 1992), a few analyses were undertaken involving the calculation of a total score (Elley, 1992, 1994).

From this discussion, the following questions arise.

1.In the context of the Reading Comprehension Study, is it appropriate to calculate a total score based on all items which were designed originally to measure different subskills in reading?

Is it possible to subsume the processes involved in (a) following the ideas of a passage, (b) finding answers that are explicitly stated in the text, (c) recognizing implied meaning, and (d) identifying a writer's purpose, under a term such as Reasoning, as proposed by Thorndike (1973a), or are they so different that they should be considered separately?
2.With respect to the Reading Literacy Study, is it appropriate to combine Narrative, Expository and Document domain scores into a construct of reading achievement or should the domain scores be used as separate outcome measures?

In other words, is there an underlying common factor or trait that allows the calculation of a total score or are the processes involved

in the reading of Narrative, Expository and Document material different to the extent that would render the calculation of a total score inappropriate?

In order to examine these questions, confirmatory factor analysis (CFA) was performed to evaluate the different hypothetical models of the structure underlying the tests which were based on the assumptions

involved in the design of the reading tests.
CFA was also used to address an additional question since Hambleton and Kanjee (1994) have pointed out that the examination of factor structures underlying tests that have been translated into different languages can be employed to determine whether or not tests were equivalent across countries.

3.Are similar results obtained across countries from an examination of the different factor structures of the Reading Comprehension and Reading Literacy tests?
Does the relative appropriateness of the factor structures underlying the processes involved in reading comprehension as assessed by the Reading Comprehension and Reading Literacy tests vary depending on the language of the test and the context of the different countries or is it similar across the eight countries that participated in both studies? Similar factor structures underlying these two tests, which were administered cross nationally, would provide evidence that problems associated with translation or culturally specific contexts are less threatening to the validity of international comparisons in practice, than they are commonly argued to be, without evidence and merely from a theoretical perspective (Hambleton and Kanjee, 1994).

Confirmatory factor analysis

Confirmatory factor analysis (CFA) is a factor analytic technique which is employed where a researcher makes certain assumptions with respect to the factorial structure underlying a test or a battery of tests and intends to evaluate these assumptions in the light of the evidence gathered (Jöreskog and Sörbom, 1993). The technique is confirmatory rather than exploratory in nature since it is used to examine how well a model which has been developed prior to the analysis fits the data. For that reason, Marsh (1991) labelled such models a priori models and distinguished them from a posteriori models, which take into consideration results from previous analyses, are changed accordingly, and then reestimated.
Hierarchical confirmatory factor analysis (HCFA) is employed for the same purpose and in the same context as CFA but incorporates an assumption of the hierarchical ordering of factors in first- or lower-order factors and second- or higher-order factors. In a hierarchical factor model (HO), items or observed variables are assigned to first-order factors which, in turn, are assigned to one or more higher-order factors.
However, Gustafsson and Balke (1993) have pointed out certain problems with respect to this approach of higher-order modelling. The authors emphasized that HO models might be misleading in that higher-order factors were only indirectly linked to the individual items through first-order factors. This, it was argued, might not be appropriate in that "the most important difference between higher- and lower-order factors is that the former affect a wider range of observed variables than do the latter" (Gustafsson and Balke, 1993, p. 415). Therefore,

Gustafsson and Balke (1993) suggested that nested factor models (NFs), in which more general or higher-order factors as well as lower-order factors are directly linked to observed variables, but with the higher-order factors set orthogonal to the lower-order factors, might be more appropriate. Gustafsson and Balke (1993) noted that HOs and NFs

may be interpreted in a similar way although the structural equations involved vary slightly. Thus, in their analyses of structural issues of intelligence, Gustafsson and Balke (1993, p. 416) use NF-modelling "in close connection with HO-modelling".

More recently, Mulaik and Quartetti (1994) have undertaken analyses to compare results produced by HO-modelling and NF-modelling from which some evidence emerged to support the HO-modelling approach. The issue, however, remained unresolved and the authors suggested that further work would be necessary before more substantiated claims as regards the appropriateness of HO- or NF-modelling could be made. While some of the factor structures under examination are represented graphically as HO-models, the actual LISREL analyses of these models followed the NF design.

As described above, the design of both the Reading Comprehension as well as the Reading Literacy tests involved certain assumptions regarding the assigning of individual items to certain categories depending either on the types of processes involved in answering the items or on certain types of reading materials. In addition, assumptions were made regarding the distinctness of these categories and the appropriateness of the use of respective sub-scores. In order to shed further light on the issue of the structure underlying the processes involved in reading comprehension, CAF and HCFA were applied to the reading test data of the Reading Comprehension and Reading Literacy Studies using LISREL8 for Windows (Jöreskog and Sörbom, 1993). It was of particular importance to examine the hypothesized models by replicating the same analyses for the eight countries involved in both studies. Thus, the focus of the analyses was to obtain an indication of the universal appropriateness of the models across countries rather than to develop a best-fitting model for a particular country.

As is explained more thoroughly below, in some of the models presented and subsequently analyzed in this paper the existence of a general factor underlying the processes involved in reading comprehension at the 14-year-old level was assumed. In other words, although there may be several traits involved when reading with understanding, it was hypothesized that the different traits or characteristics measured by the reading test items operated in unison and did not measure traits or characteristics which acted independently of or against one another. If results of the analyses showed that the available data did not contradict this assumption, then analyses involving IRT methods could be undertaken.

Hence, the purpose of the CFA analyses reported first for the Reading Comprehension, and secondly for the Reading Literacy Study is to (a) describe the different hypothesized models, (b) evaluate the models

using item data of the eight countries that participated in both studies, (c) determine which model best reflects the data, and (d) determine the appropriateness of the assumption of one major factor underlying the processes involved in reading comprehension at the 14-year-old level.

Figure 1Hypothesized factor structures of Reading Comprehension test items

Confirmatory factor analysis of Reading Comprehension test

Figure 1 illustrates three models of hypothesized factor structures underlying the Reading Comprehension test items which have been labelled RC1 to RC3 for easier reference.

In the graphical representation of the models, observed variables, that is the items given in the tests, are displayed in rectangular boxes while latent variables are shown in boxes with rounded corners. Since some models involve first- as well as higher-order factors, boxes with rounded corners are shaded differently whereby the lighter shading indicates first-order factors while the darker shading refers to higher-order factors. It should be noted that the arrangement of observed and latent variables and the arrows does not represent assumptions of a formative flow of influence but merely the reflective assigning of variables to factors.

In Figures 1 and 2, items number 1, 2, 25, 26 and the last item in each test were taken to represent the general idea of CFA in which individual items are assigned to hypothesized factors without drawing all item-factor relationships for reasons of parsimony. These items were selected since they were actually assigned to the four specific factors in the Reading Comprehension test and the three specific factors in the Reading Literacy test. The numbers, however, only indicate the item's position in the test and not that these were the items common to both tests.

RC1 presents a basic factor model in which observed items are assigned to only one single-order factor, labelled "Reasoning" (Thorndike, 1973a). This model was tested in two different ways using the Reading Comprehension item data. First, calculations were undertaken including all 52 test items. Secondly, the model was tested using those 42 items that had been assigned to reading subcategories in order to have a more appropriate comparison with the subsequent two models which also included only these 42 items.

RC2 illustrates a four single-order factor solution in which 42 out of the 52 items in the Reading Comprehension test are assigned to the four factors that are equivalent to the subcategories of: (a) following the ordering of ideas in a paragraph (Pasorder); (b) finding answers that are explicitly stated in the text (Explicit); (c) recognizing implied meaning (Implicit); and (d) identifying a writer's purpose (Evauthor)

in accordance with reading experts who had been consulted during the study. It should also be noted that each item is assigned to one and only one factor and that, as for all the Reading Comprehension factor structure models presented here, the errors of measurement are considered to be uncorrelated.

RC3 hypothesizes a hierarchical factor structure whereby 42 observed measures are assigned to the four first-order factors of reading processes, namely Pasorder, Explicit, Implicit and Evauthor which, in turn, are nested within a general higher-order factor, namely Reasoning. Further assumptions in the model, as raised above, are that while the first-order factors are allowed to be intercorrelated (as indicated by the two-way arrows), the covariance between first-order factors and the higher-order factor is set to zero. In other words, a hierarchical structure of the processes involved in reading at the 14-year-old level is assumed in which four intercorrelated single-order factors are nested within a more general, unrelated, higher-order factor.

The confirmatory factor analyses of the data were undertaken using matrices of tetrachoric correlations between items. Tetrachoric correlations were considered more appropriate since, although the available test item data from the two studies were dichotomous variables in that students' answers were coded either correct or wrong, it was reasonable to assume an underlying normal distribution. As Wood (1985, p. 377) pointed out:

> A very common assumption made in test theory is that the ability to answer a particular item varies from very low ability to very high ability in a population of subjects. It is assumed that the subjects are distributed on this item continuum according to the normal distribution […].
>
> When these distributional assumptions are taken into consideration, several authors (Clauß and Ebner, 1989; Keeves, 1988; Lord, 1980)

recommend the use of the tetrachoric correlation as a more appropriate measure of association than other correlation coefficients. They contend that the underlying normal distribution not only operates between individuals, but also within individuals with respect to each particular item. Keats (1994) maintains that the normal distribution is not necessarily the most appropriate and that the development of Strong Test Theory will inevitably occur during coming decades.

Once the tetrachoric correlation matrices were generated from the raw data using PRELIS (Jöreskog and Sörbom, 1993), LISREL8 for Windows (Jöreskog and Sörbom, 1993) was employed for analysis.

The results for the confirmatory factor analysis of the hypothesized factor-structure models underlying the Reading Comprehension test items are given in Table 1. From left to right, information is given on the particular model tested, its factor structure, the country with which data the model was tested, goodness-of-fit and adjusted goodness-of-fit

indices, the degrees of freedom (df), chi-square (c2), the c2/df ratio and the Bentler Bonett Index (BBI).
The information on the null models provided in Table 2 is only relevant in so far as the c2 values were used to compute the BBI. In null models, by definition, each item is assigned to one factor. In other words, each item is assumed to be independent of the other items and to represent a separate factor. Hence, the number of factors in a null model coincides with the number of items. Since RC1 was tested using the 42 as well as the 52 item battery, two null models were calculated for each country taking into account the different number of items.
Marsh (1991) emphasized the usefulness of other indices to evaluate higher-order factor models such as the higher-order Tucker Lewis index (HTLI) or the higher-order relative noncentrality index (HRNI). However, a decision was made to use c2/df and the BBI since the majority of models for the Reading Comprehension as well as the Reading Literacy Study represented first-order structures for which no higher-order indices could be obtained, thus making comparisons between the different models harder. Furthermore, it was argued that the range of indices provided would give a sound indication of the general fit of a model.
Byrne (1989) noted that values smaller than 2.00 for the c2/df ratio and values greater than .90 for the BBI represented an acceptable fit of the model to the observed data. It is evident that only a few of the models fitted the data well. However, it should be noted that all indices employed are influenced by the sample size which is given in Table 2. Only the three analyses for the United States fell within the recommended limits which might be a consequence of the test being developed in the United States. Despite the low level of fit of the models to the data, the analyses presented in this paper allowed a systematic comparison of the hypothesized models and a rigorous assessment of their relative appropriateness, in the absence of better theories on the factor structure of reading processes of 14-year-old students for the two studies.

Table 1Results for confirmatory factor analysis of Reading Comprehension test

Table 2 c2 for null models of Reading Comprehension test items

Notes:
a)FO-First-order factor
b) HO -Higher-order factor f) No. of items in analysis. RC1 was tested using (a) all 52 test items except for Belgium (Fr.) where item
c)Rank order (1-4) of this model (RC1 - RC3) for a country considering c2/df.

d)BBI-Bentler Bonett index: 30 had to be omitted due to coding error,

and (b) with the 42 items assigned to subskills by Thorndike (1973b). Models RC2 and RC3 were tested using 42 items assigned to subskills by Thorndike (1973b).
g)Did not converge.
h)No. of factors = No. of items for null model.
i)No. of students who answered at least three items
e)Rank order (1-4) of this model (RC1 - RC4) for a country considering BBI.overall

In order to evaluate which of the hypothesized models fitted the data best, two columns giving the rank order of each model for the respective country were added. The ninth column gives the rank according to the c2/df while in the last column the rank according to the BBI is noted. The ranking procedure may be illustrated using Belgium French as an example. Consider the c2/df ratios for the analyses that were undertaken using the Belgium French data set. Rank number one is assigned to the analysis which shows the lowest value (1.26 for RC3), hence indicating the relative best fit. RC1 using all 52 items yields the second highest value (1.32) and is, therefore, assigned rank two. RC1 involving 42 items is ranked third (1.37) while RC2, which assumes four first-order factors, has the highest c2/df ratio (1.43) indicating the worst fit and is thus ranked fourth. Likewise, the values for the BBI of the different models for Belgium French are ranked with the value closest to one indicating the best fitting model and the lowest value indicating the worst fit. When considering the BBI, the rank ordering of models for Belgium French is slightly different. While rank one is still occupied by RC3 (.56), RC1 with 42 items is now on rank two (.49) while RC2 (.47) is on rank three followed by RC1 with 52 items on rank four (.46).
In this way, ranks were assigned to each country according to the size of the c2/df ratio and the size of the BBI. A mean rank was then calculated for the two columns. As can be seen in Table 1, there is clear evidence that RC3 with a hierarchical factor structure of four first-order and one higher-order factor is the model with the best relative fit of the models under examination. For both the c2/df ratio as well as the BBI rank ordering, this model occupies the first rank in all countries. RC1, involving 42 items, is assigned the overall second rank. With respect to rank three and four, however, the ranking according to the two different indices does not coincide. While for the c2/df ratio RC1 with 52 items occupies rank three and RC2 rank four, the ordering is reversed when considering the BBI.

Confirmatory factor analysis of Reading Literacy test
Figure 2 presents the four models of the factor structures that were hypothesized to underlie the Reading Literacy test. Assumptions common to the four models include that: (a) they involve all 89 items; (b) each item may load on only one first-order factor; and (c) the items are assumed to be uncorrelated.
RL1 represents a first-order structure in which each of the 89 items is assigned to one of the three domains as defined in the Reading Literacy

Study, namely, Narrative, Expository or Document. As stated earlier, the items in the graphic model illustration serve only as an example. However, their numbers refer to their sequential position in the test and correspond to the actual attribution of these items to domains.
In RL2, all items load on one general first-order factor, namely Reasoning. Thus, it is structured in the same way as RC1 for the Reading Comprehension test. In this way, it is possible to examine the same model using data from the two studies and to compare the relative appropriateness of this model.
An underlying higher-order structure is presented in RL3 where each item is first assigned to one of the three domains as first-order

factors which, in turn, load on a general higher-order factor, Reasoning. This is a similar structuring to RC3 presented in the previous section whereby the first-order factors are allowed to be intercorrelated (as indicated by the two-way arrows) while the covariance between first-order factors and the higher-order factor, Reasoning, is set to zero. In other words, a hierarchical structure of the processes involved in reading is assumed in which three

Figure 2Hypothesized factor structures of Reading Literacy test items

Table 3Results for confirmatory factor analysis of Reading Literacy test

Table 4 c2 for null model of Reading Literacy item analysis

Notes:
a) - e)See notes Tables 1/2.
f)All Reading Literacy test items included in analysis.h)No. of factors=No. of items for null model.
g)No fit indices provided due to perfect fit of model for Finland.i) Students answering at least three items overall.

intercorrelated single-order factors are nested within a more general, unrelated or orthogonal higher-order factor.
Likewise, RL4 illustrates a higher-order structure with the alteration that the two first-order factors from RL3 are collapsed into one. From previous classical item analyses, which, for reasons of space, cannot be described in more detail here, evidence had emerged to support the combining of the Narrative and the Expository domain.
Therefore, the opportunity was taken to test whether a model assuming two correlated but separate domains of Narrative and Expository as first-order factors (RL3) would be more appropriate than a model in which the Narrative and Expository first-order factors were combined into one (RL4).
The four models were analysed using the item data from the Reading Literacy Study for the eight countries and results are presented in

Table 3. For an explanation of the headings in this table and the indices used, the reader is referred to comments regarding Table 1 in the previous section. Again, information on the null model for each country is provided (see Table 4) since, as explained earlier, information on the Chi-square values of the null models was necessary to calculate the BBI.

Except for Finland, where all models show a perfect fit, results indicate that none of the models can be considered to represent a good fit to the data. However, in the absence of alternative theoretical structures, an attempt is made to assess the relative appropriateness of the hypothesized models. Again, it should be noted that the indices employed are highly dependent on sample size, and the lack of fit may be attributable to the large sizes of the samples as is recorded in Table 4.

An explanation of the ranking procedure to assess the relative appropriateness of the hypothesized models is provided above and need not be repeated here. An examination of the mean ranks calculated across countries reveals that the model that best fits the data is RL3 which assumes a hierarchical factor structure where the three domains are first-order factors nested beneath the general factor, Reasoning. Rank two is occupied by RL4 which differs from RL3 only in that the two first-order factors Narrative and Expository are combined. This rank order of models RL3 and RL4 shows that RL3 fits the data slightly better across countries than RL4. Hence, processes involved in the reading of Expository and Narrative materials do not seem to be sufficiently close in kind to warrant a combining of the two domains. Rather, it appears to be more appropriate to consider these two domains as correlated but separate factors. It has to be kept in mind, however, that the indices for these two models presented in Table 3 only differ in the second decimal place, if they differ at all, indicating how close the models are in terms of their fit to the data. Nevertheless, where there is a slightly better index, this favours RL3.

Ranks three and four are occupied by RL1 and RL2 respectively. It is of some interest to note that RL1 with the three domains as first-order factors is a better solution than RL2 which assumes one general underlying first-order factor. In other words, a three first-order factor structure which takes into consideration the different types of reading materials employed in the Reading Literacy test represents the data better than a single first-order factor structure.

In this section, different models of the structure underlying the reading tests in the Reading Comprehension and Reading Literacy Studies were presented and evaluated using data from the eight countries that participated in both studies. Results of the confirmatory factor analyses of the models using the available data showed that none of the proposed factor structures represented a highly satisfactory model which fitted the data well across countries. Hence, further work is needed to develop appropriate indices of goodness of fit that are fully

independent of the sizes of the samples employed to establish whether the observed lack of fit is associated with the models employed or the size of the samples. Nevertheless, it may also be necessary to advance more suitable models of the structure underlying the processes involved in reading at the 14-year-old level, or to develop test items that measure subskills or processes involved in the reading comprehension of different types of reading materials more effectively.

However, an evaluation of the relative appropriateness of the different hypothesized models revealed a high level of consistency across countries. For both studies, a hierarchical factor model in which several correlated first-order factors were nested under a general, uncorrelated or orthogonal higher-order factor was the most appropriate. In the most appropriate model for the Reading Comprehension Study, items were assigned to four correlated single-order factors according to the different reading skills which, in turn, were assigned to a general, uncorrelated higher-order factor, Reasoning. In the most appropriate model for the Reading Literacy Study, each item loaded on one of the three correlated first-order factors defined by different types of reading materials which, in turn, loaded on one general higher-order factor.

Finally, in the analyses undertaken, no evidence emerged to reject the assumption of the existence of one general higher-order factor underlying the processes involved in reading at the 14-year-old level and there was clear evidence of orthogonal lower-order factors nested beneath. This evidence would suggest that it is appropriate and consistent to present results in both studies in terms of a total score as well as the appropriate subscores.

Rasch scaling of Reading Comprehension and Reading Literacy item data

Rasch scaling of the Reading Comprehension and Reading Literacy item data was undertaken to (a) examine whether or not it was possible to scale the items from the two reading tests using the one-parameter Rasch model, which assumes a common underlying trait; (b) identify items which do not fit the scale well across countries; and (c) check whether or not the poor fit could stem from problems that are associated with translation or the culturally specific context of an item. These issues had to be considered before an attempt could be made to equate the reading performance of students over the 20-year-period. The results of the confirmatory factor analyses had indicated that the

items of the Reading Comprehension and Reading Literacy tests conformed to a structure which assumed one general factor underlying the processes involved in reading beyond the decoding stage. Furthermore, results of the analyses suggested that the model which assumed several first-order factors nested beneath a general higher-order factor was the most appropriate both for the Reading Comprehension and Reading Literacy item data. This evidence suggested that the calculation of a total score was consistent with the data in that the items in the two tests could be considered to measure the general underlying

higher-order factor.

The concept of unidimensionality of the characteristic or trait under investigation is one of the major assumptions made by Rasch (1960) to develop a theory of item response in which respondents are located along a continuum of the latent trait under investigation. The term latent indicates that the underlying trait, which might be an ability or aptitude, may not be directly measured. Therefore, tests are employed to obtain information which allows the estimation of a respondent's position on the continuum of the underlying trait (Hambleton and Swaminathan, 1985).

While it is frequently emphasized (Keeves, 1992; Weiss and Yoes, 1991) that most psychological and educational tests focus on a particular aspect and are, therefore, likely to comply with the assumption of a single latent trait, the concept of unidimensionality as such is vague and few criteria have been established to characterize it. At a first glance, the idea of unidimensionality appears to oversimplify real-life processes especially in the areas of psychology and education where the general assumption is that many different factors are operating. Bejar (1983), however, has stressed that for unidimensionality to apply it is not necessary for only one single underlying process to be operating but it is essential that, where different processes are involved, they operate in unison.

Further to the one-parameter model put forward by Rasch (1960), other models based on Item Response Theory (IRT) have been developed. While the two-parameter model (Birnbaum, 1968) adjusts for differences in item difficulties and differences in slopes, the three-parameter model (Lord, 1952), in addition, incorporates a correction for guessing as a characteristic of an item and not of a person. Over and above adjusting for differences in item difficulty and guessing, the four-parameter model (McDonald, 1967) also takes into account the possibility of carelessness. Carelessness, in this context, refers to the possibility of the item difficulty having an effect on a person's tendency to respond whereby very able students may become careless with easy items and therefore provide wrong answers.

In order to compare the different approaches, Sontag (1983) evaluated the results obtained by applying the different models. These results indicated that the one-parameter model provided more stable estimates than the two- and three-parameter models when comparing achievement levels of different age groups with multiple-choice test items of the kind being examined in this paper.

Where the basic assumptions underlying IRT models are met, items and respondents may be measured conjointly on the same scale. The relative relationship between a person's level of ability and the item difficulty level is associated with the probability of a person answering an item correctly. In addition, IRT produces: (a) an item estimate which is independent of the sample of people taking a particular test; and (b) an estimate of a person's score which is independent of the items taken, provided that the assessed phenomenon involves processes that conform to an underlying or latent trait (Keeves, 1992; Stocking, 1994). Furthermore, Keeves (1992) pointed out

that models based on the assumption of an unobservable latent trait that can be expressed as a logistic function allow the vertical as well as horizontal equating of that trait. In other words, the use of IRT in the scaling of tests would allow comparisons across age groups as well as comparisons of the same age group over time.

The scaling of item data from the Reading Comprehension and Reading Literacy Studies, was undertaken with the complete set of items using Quest (Adams and Khoo, 1993), a program for test analysis, which employs a Rasch or one-parameter logistic model (Rasch, 1960). Results of the analyses are presented in Appendix A to Appendix D. While Appendices A and C show the thresholds (item difficulty) for the Reading Comprehension Study and Reading Literacy Study, Appendices B and D present the infit mean square values (item discrimination) for the two studies.

The item characteristics in the form of thresholds and infit mean square values were then examined to identify misfitting items. An item was considered not to fit the scale where the threshold value was outside ±0.70 of the mean threshold value in more than three countries (Wright and Douglas, 1975). In addition, infit mean square values for items outside a range from 0.8 to 1.2 were considered acceptable. This range would appear to lie between the more rigorous values employed by Keeves and Schleicher (1992), who judged a range from 0.88 to 1.13 to be appropriate, and Adams and Khoo (1993), who recommended from extensive experience a range from 0.75 to 1.30. In this way, seven items of the Reading Comprehension item data and eight items of the Reading Literacy item data were excluded from the Rasch scaling of items for the equating process.

Table 5 presents the Rasch estimates of item thresholds for misfitting items in the Reading Comprehension test and Table 6 shows the Rasch estimates of item thresholds for misfitting items in the Reading Literacy test. The infit mean square values are not presented since the values lay within the specified acceptable range with the exception of only three values across all countries and all items in the two studies. Thus, items were omitted on the basis of the spread in threshold values which indicated that items varied considerably in their relative difficulty levels across the eight countries. In order to investigate the extent to which problems with translation or culturally specific content might have contributed to the spread in threshold values, the items and their corresponding values in Tables 5 and 6 were examined in more detail.

In Table 5, two values stand out clearly: The threshold value for item 30 in Belgium French is 5.16 and the threshold value for item 38 is 3.12 in the United States. Results of the classical item analysis which was undertaken prior to the analyses presented in this paper showed that in both cases a majority of students had not chosen the correct answer but one of the distractors. While in the case of item 30 this might have been a result of a translation error in the Belgium French reading test, the high number of incorrect answers to item 38 in the

United States might be a result of the fact that students had perform a minor calculation involving the use of kilogram as a unit in order to arrive at the right answer. The metric measure might have confused students somewhat. The fact that a relatively high positive threshold value, indicating a high degree of difficulty, is also recorded for New Zealand (1.42) suggests that the use of the metric measurements for 14-year-old students in 1970/71 was more difficult for students from New Zealand and the United States than for their European counterparts. It the impact of these high threshold values on the size of the mean threshold value across countries that item 30 and item 38 appear not to fit the scale. If the two values were excluded from the calculation, the remaining threshold values would not lie outside the acceptable range in more than three countries.

The highest number of threshold values that lie outside the acceptable range in Table 5 are recorded for New Zealand and the United States. Moreover, the threshold were located towards the more difficult end of the item continuum. This seems to indicate that the items were

relatively easier in the translated version than in the language in which the tests were developed, namely English.

For the Reading Literacy Study, this picture is reversed, with the fewest misfitting items recorded for New Zealand and the United States in Table 6. Here, it is interesting to note that four of the eight misfitting items (items 6, 7, 9, and 10) are based on one passage, namely the filling in of a travel card using the information provided in the text. The high negative item thresholds indicate that these items were very easy in relation to the other items in the test in all countries. Across countries, however, values indicate that these items were relatively harder for students in the Netherlands and Sweden and extremely easy for students in Belgium French. Indeed, all students used in calibration for Belgium French answered item 7 in the Reading Literacy test correctly.

Table 5Rasch estimates of item thresholds for misfitting items - Reading Comprehension Study

Table 6Rasch estimates of item thresholds for misfitting items - Reading Literacy Study

Note:
a)Number of misfitting items per country.

Since all three countries are located in Europe where travel experiences would be similar, there appears to be no reason why students in Belgium French should be more familiar with the filling in of a travel card than students in the Netherlands and Sweden.

From this examination of threshold values for the items which had been identified to fit the scale poorly because values were outside the acceptable range in more than three countries, no systematic reason emerged as to the poor fit of the items. An examination of the individual items suggested that the poor fit of some items could stem from translation errors or varying degrees of student familiarity with certain units of measurement. However, no systematic pattern as to the reason for the spread of item threshold values across countries emerged.

Summary and Conclusion

The analyses presented in this paper may be summarized as follows.

1.Confirmatory factor analysis revealed that none of the hypothesized models of the factor structure underlying the Reading Comprehension or Reading Literacy tests represented a good fit to the data, but this could well be a consequence of the dependence of the indices employed on the sample sizes which were very large.

2.An appraisal of the relative appropriateness of the hypothesized models, however, showed that:

(a)the model which best reflected the Reading Comprehension data was a hierarchical model with one general higher-order factor, namely Reasoning, under which four correlated first-order factors, namely the ordering of ideas in a paragraph, finding answers explicitly stated in the text, recognizing implied statements and identifying the author's purpose, were nested.

(b)the model which best reflected the Reading Literacy data was a hierarchical factor model which had one general higher-order factor, namely Reasoning, under which three correlated first-order factors, namely the Narrative, Expository and Document domain were nested.

3.Similar results were obtained across countries from an examination of the different factor structures. While the different factor structures showed varying degrees of fit from country to country, a high degree of similarity emerged with respect to the best fitting factor structure across countries.

4.From the confirmatory factor analysis no evidence emerged against the assumption of a general latent trait underlying the processes involved in reading comprehension.

5.With very few exceptions, the items of both the Reading Comprehension and the Reading Literacy tests could be scaled using the one-parameter Rasch model.

6.The issues of translation and context of items that have been repeatedly emphasized from a "data-free" perspective (Hambleton and Kanjee, 1994) could be examined effectively in terms of differences in threshold levels across countries. Nearly all items from both tests showed similar threshold levels across countries. From an examination of those items which varied considerably in their difficulty levels across countries, no consistent pattern emerged as to the reasons for the misfit.

7.Results of the (a) Rasch scaling of the item data from the two studies; and (b) examination of factors structures from which similar results had been obtained across countries, provided supportive evidence of the conclusion reported by Thorndike (1973b) that translation problems could be overcome.

… it [i.e. the consistency of item characteristics] does indicate a substantial core of similarity in the Reading Comprehension task as one goes from country to country and from language to language. (Thorndike 1973b, p. 164)

References

Adams, R.J. and Khoo, S.K. 1993 Quest - The interactive test analysis system. ACER, Hawthorn, Vic.
Bejar, I.I. 1983 Achievement testing: Recent advances. Sage university paper series on quantitative applications in the social sciences. Series No. 07-036. Sage Publications, Beverly Hills, California.
Birnbaum, A. 1968 Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick Statistical theories of mental test scores. Addison-Wesley, Reading, M.A.
Byrne, B.M. 1989 A primer of LISREL. Basic applications and programming for confirmatory factor analytic models. Springer-Verlag, New York.
Clark, M.L. 1973 Hierarchical structure of comprehension skills. Volume 2. Australian Council for Educational Research, Hawthorn, Victoria.
Clauß, G. and Ebner, H. 1989 Statistik für Soziologen, Pädagogen, Psychologen und Mediziner. Band 1: Grundlagen. Verlag Harri Deutsch, Thun und Franfurt am Main.
Davis, F.B. 1944 Fundamental factors of comprehension in reading. Psychometrika, 9, 185-197. op. cit. in R.L. Thorndike 1973b.
Davis, F.B. 1968 Research in comprehension in reading. Reading Research Quarterly, 3, 499-545.
Elley, W.B. 1992 How in the world do students read? The International Association for the Evaluation of Educational Achievement, Hamburg.
Elley, W.B. (ed.) 1994 The IEA study of reading literacy: Achievement and instruction in thirty-two school systems. Pergamon Press, Oxford.
Gustafsson, J. and Balke, G. 1993 General and specific abilities as predictors of school achievement. Multivariate Behavioral Research, 28, 407-434.
Guthrie, J.T. 1987 Indicators of reading education. Center for Policy Research in Education. Eagleton Institute of Politics, Rutgers, The State University of New Jersey, New Brunswick, NJ. ERIC database, ref. ED291083

Hambleton, R.K. and Kanjee, A. 1994 Test and attitudes scales, Translation of. In T. Husén and T.N. Postlethwaite The international encyclopedia of education. 2nd ed. Pergamon Press, Oxford, pp.

6328-6334.
Hambleton, R.K. and Swaminathan, H. 1985 Item response theory. Principles and applications. Kluwer Nijhoff, Boston, Massachusetts.
Jöreskog, K.G. and Sörbom, D. 1993 LISREL 8: Structural equation modeling with the SIMPLIS command language. Lawrence Erlbaum Associates, Hillsdale, N.J.
Keats, J. A. 1994 Classical test theory. In T. Husén and T.N. Postlethwaite The international encyclopedia of education. 2nd ed. Pergamon Press, Oxford, pp. 785-792.
Keeves, J.P. 1988 Correlational procedures. In J.P. Keeves (ed.) Educational research, methodology and measurement: An international handbook. Pergamon Press, Oxford.
Keeves, J.P. 1992 Scaling achievement test scores. In J.P. Keeves The IEA technical handbook. International Association for the Evaluation of Educational Achievement (IEA), The Hague. pp. 107-125.
Keeves, J.P. and Schleicher, A. 1992 Changes in science achievement: 1970-84. In J.P. Keeves (ed.) The IEA study of science III: Changes in science education and achievement: 1970 to 1984. Pergamon Press, Oxford, pp. 263-290.
Lord, F.M. 1952 A theory of test scores. Psychometric Monograph No. 7.
Lord, F.M. 1980 Applications of item response theory to practical testing problems. Erlbaum, Hillsdale, NJ.
Ludlow, L.H. and Hillocks, G. 1985 Psychometric considerations in the analysis of reading skill hierarchies. Journal of Experimental Education, 54 (1), 15-21.
Lundberg, I. and Lynnakylä, P. 1993 Teaching reading around the world. The International Association for the Evaluation of Educational Achievement, Hamburg.
Marsh, H.W. 1991 Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. Journal of Educational Psychology, 83 (2), 285-296.
McDonald, R.P. 1967 Non-linear factor analysis. Psychometric Monograph, No. 15.
Mulaik, S.A. and Quartetti, D.A. 1994 First-order or higher-order? Paper presented to the Society of Multivariate Experimental Psychology, Annual Meeting, Princeton, New Jersey, October 28. 1994.
National Assessment of Educational Progress (NAEP) 1971 Reading objectives. First assessment 1970/71.
National Assessment of Educational Progress (NAEP) 1984 Reading objectives. 1983-84 assessment.
O'Neill, O.C. 1978 A factorial analysis of reading comprehension skills at the secondary level. Unpublished Masters thesis, The Flinders University of South Australia.
Postlethwaite, T.N. and Ross, K.N. 1992 Effective schools in reading. Implications for educational planners. The International Association for the Evaluation of Educational Achievement, Hamburg.
Rasch, G. 1960 Probabilistic models for some intelligence and attainment tests. Danmarks Paedagogiske Institute, Copenhagen.
Reutzel, D.R. and Hollingsworth, P.M. 1991 Testing the distinctiveness hypothesis. Reading Research and Instruction, 30(2), 32-46.

Sontag, L. M. 1983 Vertical equating methods: A comparative study of their efficacy. PhD thesis. Teachers College, Columbia University, New York.
Spearritt, D. 1972 Identification of subskills in reading comprehension. Reading Research Quarterly, 8 (1), 92-111.
Stocking, M.L. 1994 Item response theory. In T. Husén and T.N. Postlethwaite The international encyclopedia of education.(2nd ed.). Elsevier Science, Oxford, pp. 3051-3055.
Thorndike, R.L. 1973a Reading as reasoning. Reading Research Quarterly,

2, 135-147.
Thorndike, R.L. 1973b Reading comprehension education in fifteen countries. International studies in evaluation III. Almqvist and Wiksell, Stockholm.
Weiss, D.J. and Yoes, M.E. 1991 Item response theory. In R.K. Hambleton and J.N. Zaal Advances in educational and psychological testing: Theory and Applications. Kluwer Academic Publishers, Boston. pp. 69-95.
Wood, R. 1985 Item analysis. In T. Husén and T.N. Postlethwaite The international encyclopedia of education. Pergamon Press, Oxford, pp. 376-384.
Wright, B.D. and Douglas, G.A. 1975 Best test design and self-tailored testing. Research Memorandum No. 19. Statistical Laboratory Department of Education. The University of Chicago.
Zwick, R. 1987 Assessment of the dimensionality of NAEP year 15 reading data. In A. Beaton The NAEP 1983-84 technical report. Princeton, New Jersey, pp. 245-284.

Appendix ARasch estimates of item thresholds for Reading Comprehension items - Calibration based on students who answered all items

Notes:
a)Wright and Douglas (1975).
b)Calibration based on students who attempted all items.

Appendix BRasch estimates of infit mean squares for Reading Comprehension items

10.90.90.90.91.00.90.91.0
20.90.90.91.00.91.00.90.9
31.00.90.91.00.91.01.00.9
41.01.11.11.11.01.01.01.0
51.11.11.11.01.11.11.21.1
61.01.01.11.11.11.01.01.1
71.11.01.01.01.01.01.00.9
80.90.90.90.90.90.90.90.9
91.01.01.01.01.00.91.01.0

101.01.01.01.01.01.01.00.9
111.01.01.01.01.11.01.01.1
121.11.21.11.01.01.11.11.1
130.91.01.11.01.00.91.00.9
141.01.00.91.01.01.01.00.9
151.01.01.00.91.00.90.90.9
161.01.01.00.91.00.91.00.9
171.01.11.11.10.91.11.21.2
181.11.01.01.01.21.11.11.1
191.01.01.01.01.01.01.11.0
201.10.90.91.01.01.11.21.1
211.01.00.90.91.00.91.01.0
221.01.01.01.01.01.01.01.1
230.90.91.00.90.90.90.90.9
240.90.90.90.90.90.90.90.8
250.91.11.01.11.10.91.01.0
261.21.31.21.01.21.31.11.2
271.01.01.10.91.01.00.91.0
281.00.91.00.91.00.90.91.0
291.01.00.91.01.01.01.01.0
301.01.10.90.91.00.90.91.0
311.01.01.01.01.11.01.01.0
321.01.01.00.91.01.01.01.0

331.01.01.01.01.00.91.00.9
341.01.01.01.01.01.01.01.0
350.90.91.01.01.01.00.90.9
360.90.91.00.90.91.01.01.0
371.11.01.11.11.21.11.01.1
381.01.01.01.11.01.01.01.1
391.00.91.00.90.90.90.90.9
400.91.00.90.90.91.00.90.9
411.01.01.01.11.01.21.11.2
420.80.90.90.90.90.90.80.9
431.00.90.91.10.90.91.00.9
441.11.01.01.11.01.01.11.0
451.11.01.11.11.01.00.90.9
461.00.91.00.91.00.90.90.9
471.11.11.01.11.01.01.01.0
480.90.91.00.90.90.90.90.9
491.11.11.01.11.01.11.01.1
501.01.01.11.11.11.01.11.0
511.11.11.21.21.11.31.11.1
520.91.00.91.00.90.90.91.0

Appendix CRasch estimates of item thresholds for Reading Literacy items - Calibration based on students who answered all items

Note:
a)These values were estimated from the "Table of Measures on complete test of items" of the Quest program (Adams and Khoo, 1993) since all students used in calibration answered these items correctly in Belgium French.

Appendix CRasch estimates of item thresholds for Reading Literacy items - Calibration based on students who answered all items (continued)

Notes:
aWright and Douglas (1975).
bCalibration based on students who attempted all items.

Appendix DRasch estimates of infit mean squares - Reading Literacy Study

11.11.01.01.01.01.01.01.0461.01.01.01.01.11.11.01.0
21.01.01.01.01.01.01.00.9471.01.01.01.01.11.01.00.9
31.01.11.01.01.10.91.01.0481.01.01.01.01.01.01.00.9
41.11.01.01.01.01.01.01.0491.01.01.11.00.90.90.90.9
50.90.91.01.01.01.01.01.1501.00.90.91.01.01.01.00.9
61.01.00.91.01.01.01.01.0511.01.01.01.01.01.01.00.9
7a1.00.91.00.90.91.11.0521.11.11.01.01.01.01.01.1
8a1.01.01.01.01.01.01.01.0531.01.01.01.11.01.11.01.1
91.01.11.01.01.01.01.11.1541.01.11.01.01.01.21.11.1
101.01.21.01.01.11.01.11.1551.01.01.01.01.01.01.01.0
111.01.01.21.01.11.11.01.2561.01.01.00.91.01.01.01.0
121.01.01.01.01.11.11.01.0571.11.01.01.11.01.01.11.1
131.01.01.01.01.01.01.01.1581.01.01.01.01.01.01.01.0
141.01.01.01.01.01.00.91.0591.11.11.01.11.11.11.01.1
151.11.01.01.01.11.01.01.1601.01.11.01.01.01.11.01.0
161.11.01.01.01.01.01.01.0611.01.01.01.01.01.01.01.1
171.01.01.01.01.01.01.01.0621.01.01.01.11.01.11.11.1
181.11.01.11.01.21.11.01.1631.01.11.01.01.11.01.01.1
191.01.31.01.01.01.01.01.0641.01.11.01.01.01.11.11.1
201.01.01.01.01.01.01.00.91.0651.01.01.01.01.01.01.01.1

211.01.01.01.01.01.01.01.0661.01.01.01.01.01.01.01.0
221.11.11.11.01.11.21.11.2671.01.01.01.01.01.01.00.9
231.01.11.11.11.11.21.11.1681.00.91.01.00.90.91.01.0
241.10.91.11.11.01.01.01.0691.00.90.90.90.91.01.01.0
251.11.11.01.01.01.11.01.1700.91.01.00.90.90.91.00.8
260.91.01.01.01.01.01.01.0711.00.91.01.01.01.01.01.0
270.90.90.90.90.90.90.90.9720.91.00.91.00.90.91.00.9
281.01.01.01.00.90.90.91.0730.90.91.00.91.00.91.00.9
291.00.91.01.01.00.90.91.0740.91.01.00.90.90.91.01.0
301.01.01.01.11.01.01.01.0750.91.01.01.00.90.90.90.9
311.00.90.90.90.90.90.90.9761.01.01.01.01.00.91.00.9

321.11.01.00.91.01.00.91.0771.01.01.01.00.91.00.91.0
331.01.00.90.90.90.90.90.9780.90.90.90.90.90.90.90.9
340.90.90.90.90.90.90.90.9791.00.91.01.00.90.90.90.9
351.01.11.01.11.11.11.01.1800.91.01.01.01.01.01.01.0
360.90.90.91.00.90.91.00.9811.01.01.01.01.01.01.00.9
371.11.11.21.11.21.01.11.0821.01.11.11.00.91.11.00.9
381.11.01.11.01.01.11.21.0830.90.91.01.01.00.90.90.9
391.01.11.11.11.11.01.11.0841.00.91.01.01.01.00.90.9
400.90.91.00.91.00.90.90.9850.91.01.00.91.00.90.90.9
410.90.90.90.91.01.01.01.0860.91.01.01.01.00.91.00.9
421.01.11.01.01.11.01.11.0870.91.01.01.00.91.01.00.9
431.01.01.01.01.01.11.11.0881.01.01.01.01.00.91.00.9
441.01.01.01.11.01.01.01.0890.90.91.01.01.00.90.91.0
451.21.01.01.11.11.11.11.0

Note:
a)No values were obtained since all students used in calibration answered these items correctly in Belgium French.