Validation of the International Reading Literacy Test

Evidence from Dutch

Mieke van Diepen, Ludo Verhoeven, Cor Aarnoutse
and Anna M.T. Bosman
Radboud University Nijmegen

In 2001, the International Association for the Evaluation of Educational Achievement (IEA) conducted a comparative study of reading literacy (PIRLS 2001). A reading comprehension assessment instrument was developed and translated into the languages of 35 participating countries for this purpose. After field testing of the instrument, the final version of the Reading Literacy Test (RLT) was established. In two studies, the validity of the Dutch version of the RLT was examined. In the first study, comparison of the linguistic characteristics of the Dutch and English versions of the test showed the Dutch passages and items to contain both a greater number of words and longer words than the English passages and items. However, the use of more and longer words did not produce a higher level of complexity with respect to content, sentence structure, text structure, or test items as judged by a panel of bilingual experts. While the Dutch children had to read more and longer words than the English children, moreover, they had no problems finishing the test within the allocated amount of time. In the second study, the possible impact of the changes made after the field testing of the RLT was examined. The omission of passages and the modification or omission of test items were found to have no consequences for the psychometric properties of the Dutch version of the test were examined.

The International Association for the Evaluation of Educational Achievement (IEA) was founded in 1959 for the conduct of international comparative studies. The IEA is an independent association that now encompasses more than 60 countries. The aim of the organization is to improve education via the study of student achievement and the factors associated with student achievement in educational systems around the world. In 2001, the IEA conducted an international comparative study of the reading literacy of nine- and ten-year-old children in 35 countries. For this Progress in International Reading Literacy Study (henceforth,

Written Language & Literacy 10:1 (2007), 1–23.
ISSN 1387–6732 / E-ISSN 1570–6001

PIRLS), an International Reading Literacy Test was developed. Different linguistic versions of the same instrument were formulated for all of the participating countries. The focus of the present article is on the validation of the Dutch version of the Reading Literacy Test (henceforth, RLT).

Use of the same basic test across different cultural and linguistic settings does raise the problem of bias, a well-known phenomenon that has been studied by many researchers. For example Sireci (1997) states that when the items of a test have been translated, it is simply not possible to determine whether the differences in the success rates for two populations are caused by unequal difficulty of the items (i.e., bias) or unequal levels of competence. According to Bonnet et al. (2001), bias is the case when the elements of a test result in poorer results for one or several subjects irrespective of assessed competence. Similarly, Holland and Wainer (1993) state that an item is biased when equally proficient individuals from different groups do not have equal probabilities of answering the item correctly. Camilli and Shepard (1994) have simply characterized bias as "a kind of invalidity that harms one group more than another" while Uiterwijk (2001) has more recently defined bias as a systematic underestimation or overestimation of a parameter due to being part of a specific subgroup. Uiterwijk further observes that the construct validity of a test is in danger when more than one skill is needed to correctly answer the item and/or the necessary skills are not equal across different subgroups. In other words, the items constituting a reading literacy test should only assess reading literacy skills and not a student's prior knowledge or other — more or less irrelevant — skills. Many researchers, including Bonnet et al. (2001), have considered the linguistic and cultural implications of the translation of materials. De Jong and Vallen (1989) have emphasized the presence of linguistic and cultural factors as possible sources of item bias. Such linguistic factors as language and text features can affect the difficulty of a test at three levels: the word level (e.g., vocabulary, word frequency, or ambiguity), the syntactic level (e.g., structure or context), and the text level (e.g., difficult or ambiguous references). Such cultural factors as differences in prior knowledge due to different cultural backgrounds but also familiarity with the type of test and/or item format (i.e., "testwiseness") can also affect the difficulty of a test. Students who have learned to apply certain strategies for a particular type of test or are familiar with the desired type of response are clearly placed at an advantage. According to de Jong and Vallen (1989), moreover, linguistic and cultural factors can interact. Along these lines, Spilich, Vesonder, and Chiesi (1979) found the influence of cultural factors to be larger than the influence of linguistic factors. They also showed readers with greater prior knowledge at their disposal to be more capable of recalling important information from a text and producing a cohesive and complete report of the text.

In order to be able to compare the skills of individuals from many countries across the world and minimize the chances of test bias, the IEA applied very high methodological standards for the development of the PIRLS' Reading Literacy Test (PIRLS, 2000). To minimize cultural bias, all of the participating countries collaborated closely. Representatives from all of the countries helped to select the passages to be read and contributed to the development of the test items. The different countries were also asked to point out any cultural incompatibilities during the early stages of test development. It was decided not to use a passage about stars falling out of the sky, for example, because the passage was judged to be inappropriate for certain religious groups. Similarly, a question about using glue to catch mice in a friendly manner was omitted because it turned out that people in Italy use glue to catch mice in a cruel manner. With regard to the translation standards, it was attempted to have the RLT translated into the various target languages with as few changes of meaning, difficulty, and layout as possible.

A major step in the development of an unbiased test instrument is the conduct of a field test. The IEA thus undertook a large-scale field testing in 27 countries to evaluate the psychometric properties of the RLT. Those passages and items with the best measurement properties were selected and — as needed — adapted for inclusion in the final version of the test. According to Item Response Theory, observed item responses should be due to one underlying skill — which was reading literacy within the context of the present study — and item bias is present when the observed responses cannot be explained on the basis of that single underlying skill. An item-characteristic curve is used to depict the probability of an item being correctly answered on the basis of the relevant underlying skill, and three parameters were used to do this within the context of the present study: difficulty level, discrimination, and the chances of correctly guessing an answer. An item was considered biased when the differences between two groups for any of the aforementioned parameters were found to be significant. And in such a case, the item was either omitted or modified.

The high methodological standards of the IEA also affected the test administration procedure for the field test as well as for the final main test. The test administrators were instructed to not answer any questions with regard to the content of the test items, not provide additional instructions, not admit any latecomers, and not allow the students to leave the test session. They were also instructed to check that the students were following the instructions appropriately, and if they weren't, to repeat the particular instruction. After the test administration, the staffs of the national centers within the different countries processed the results. The responses to the open-ended questions were evaluated using the PIRLS scoring procedure. Scorers assigned a score of 1, 2, or 3 points to the open-ended items. They had to be conscientious, attentive to detail, knowledgeable of reading, and willing to

apply the scoring guidelines as instructed even if they disagreed with a particular definition or categorization. The scores for the open-ended items and the answers to the multiple-choice questions were next entered into the computer using special software (WinDEM) provided by the IEA. This software credited the answers on the multiple-choice questions (1 score point for a correct answer). The same software was also used to check and verify the test data.

Despite the efforts of the IEA to minimize cultural biases, the possibility of additional linguistic and/or psychometric biases following the modification of the instrument on the basis of the field results cannot be excluded. In the first study reported on here, thus, the following questions were considered. Do the linguistic characteristics of the Dutch version of the RLT differ from the linguistic characteristics of the international — English — version of the RLT? And if so, do the differences between the two versions of the test appear to place the students in the Netherlands at a relative advantage or disadvantage when compared to the students in the other participating countries? The international English version was used as the basic version for translation into the different languages. English-speaking countries adapted the international version where necessary. In order to answer the foregoing questions, the number of characters, number of words, mean word length, and mean sentence length for the Dutch and international English versions of the RLT were compared and further related to the difficulty of the two versions of the test as judged by a panel of bilingual experts.

In the second study reported on here, the psychometric properties of the Dutch version of the RLT were considered with respect to the following question: What are the consequences of the international decisions made on the basis of the field test results for the Dutch version of the RLT? The goal of the field test was to identify those eight passages of text with the best psychometric properties and, although most of the associated test items subsequently went unchanged, some of the items were nevertheless omitted or modified. This situation raises the question of whether the omission of certain passages and the omission or modification of certain items may have influenced the Dutch version of the test in terms of the internal consistency, inter-scorer reliabilities, item-country interactions, and percentages of correct responding.

Study 1: Linguistic characteristics

In the first study, several linguistic properties of the PIRLS RLT were explored. We expected the Dutch and English versions of the test to differ with respect to the number of words, the number of characters, the mean length of the words, and the mean length of the sentences used in the passages and test items. Given

that we could not find relevant studies of the linguistic differences between Dutch and English, we decided to initially compare some passages of the English novel *The Notebook* by Nicholas Sparks (1996) to the same passages from the almost literal translation of the book into Dutch by Servaas Goddijn, *Het Dagboek*. Both a greater number of words and longer words were used in the Dutch version of the book when compared to the English version. The details of this comparison are presented in Table 1, where more characters and longer words are also found to characterize the Dutch versions of two law texts for the European Community and European Bank when compared to the English versions. However, a greater number of words was used for the nonfiction law texts in English than in Dutch. People who read Dutch and English nonfiction texts on a daily basis report experiencing Dutch texts to be less concise (with British English slightly less concise than American English, in turn). However, translators also report Dutch text to contain fewer words than English texts due to the compounding of terms in Dutch but not in English (e.g., *onderwijskunde* = educational sciences; *onwikkelingslanden* = developing countries).

Table 1. Linguistic Characteristics of the Dutch and English Versions of Three Passages from a Novel and Two Law Texts

	Number of words		Number of characters		Average word length	
	Dutch	English	Dutch	English	Dutch	English
Literary	932	874	4144	3842	4.45	4.40
Informative	2641	2800	14973	14635	5.67	5.23

If the Dutch version of the PIRLS RLT contains a greater number of words, longer words on average, and longer sentences on average, then the Dutch users of the test may require more time and energy to understand the relevant texts and thus be placed at a relative disadvantage. Alternatively, the more extensive phrasing of the Dutch passages may place the Dutch users at an advantage as a greater number of characters provides more information. In addition, it is also possible that the use of more characters is associated with *more* complicated sentence and text structures but *less* complicated content, which is what we expected to find for the PIRLS RLT.

Method

Materials

To provide a valid and reliable measure of reading skill, the IEA stated that at least four hours of assessment were necessary, which amounts to eight assessment blocks (i.e., passages of text with accompanying items). In the field test, a total of

16 assessment blocks was evaluated.[1] These 16 assessment blocks were distributed across eight booklets, which thus contained two blocks for administration to a student. For each block, an average of 13 items was developed to produce between 16 and 19 points.

In PIRLS, reading literacy was defined as "the ability to understand and use those written language forms required by society and/ or valued by the individual" (Campbell et al., 2001, p.3). Young readers can read to learn, to participate in a community of readers, or for enjoyment. Given that reading literacy is directly related to the reasons for reading, the PIRLS assessment was concentrated on the two most pervasive reasons presented by young students for reading: Reading for literary experience/ enjoyment or reading to acquire and use information.

The accompanying test items assessed four comprehension processes necessary for reading literacy: (a) focus on and retrieval of explicitly stated information, (b) the drawing of straightforward inferences, (c) interpretation and integration of ideas and information, and (d) examination and evaluation of content, language, and textual elements. Both open-ended and multiple-choice items were used. The open-ended items were used to have the students formulate their own views, present their own interpretations and evaluations of the text, and explain their reasoning. They could yield 1, 2, or 3 points. A scoring guide was used to assign points to the open-ended item responses. The multiple-choice questions offered four plausible response options; only one option was correct or clearly the best response for the question. Each of the multiple-choice questions yielded 1 point when answered correctly.

Procedure

In order to compare the linguistic characteristics of the Dutch and English versions of the RLT, the total numbers of characters, words, and sentences were counted. The mean word length (i.e., total number of characters divided by the total number of words) and mean sentence length (i.e., total number of words divided by the total number of sentences) were calculated next. Analyses of variance were then performed to compare the linguistic characteristics of the Dutch and English versions of the test with Language (Dutch vs. English) as a between-subjects variable and Type of text (i.e., literary vs. informative) as a within-subjects variable. The correlations between the linguistic characteristics of the Dutch and English text passages and test items were also calculated.

A panel of six Dutch reading experts was asked to independently evaluate the complexity of the Dutch and English versions of the test. Each expert read eight randomly ordered passages in Dutch and eight randomly ordered passages in English and also answered the accompanying test items. After this, they evaluated the complexity of the content, the text structure, the sentence structure, and the test

items along a five-point-scale. To compare the Dutch and English versions of the test, t-tests were performed on the different complexity measures. Finally, the correlations between the various complexity measures both within and across the different versions of the test were explored.

Results

Linguistic characteristics

In Table 2, the linguistic characteristics of the English version and the linguistic characteristics of the Dutch version of the field-tested PIRLS RLT are presented. More specifically, the means and standard deviations for the text passages and accompanying items are presented for the eight informative texts and eight literary texts in Dutch and English.

Table 2. Linguistic characteristics of Dutch and English versions of the field-tested PIRLS RLT

Test version	Linguistic characteristic	Informative texts		Literary texts		Total	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Dutch passages	N Words	629	131	673	140	651	133
	N Characters	3043	612	2988	596	3015	585
	Word Length	4.8	0.1	4.4	0.1	4.6	0.2
	Sentence Length	12.9	1.9	12.0	2.3	12.4	2.1
English passages	N Words	618	130	641	126	629	124
	N Characters	2797	567	2755	504	2776	519
	Word Length	4.5	0.1	4.3	0.1	4.4	0.2
	Sentence Length	12.6	2.0	11.5	2.4	12.0	2.2
Dutch items	N Words	450	72	433	75	441	72
	N Characters	1982	309	1782	317	1882	320
	Word Length	4.4	0.1	4.1	0.1	4.3	0.2
English items	N Words	405	86	414	82	410	81
	N Characters	1758	299	1614	313	1686	305
	Word Length	4.4	0.5	3.9	3.8	4.1	0.5

Separate two (Type of text: literary vs. informative) by two (Language: Dutch vs. English) analyses of variance were performed on the number of words in the *passages* and the number of words in the *items*, respectively. The same analyses were also performed on the number of characters in the *passages* and the number of characters in the *items*. The main effect of Language on the number of words in the items was significant ($F(1,14) = 12.91$, $p < .01$), while the main effect of Language on the number of words in the passages was not ($F(1,14) = 1.70$, $p > .05$). Relatively more words were thus used in the Dutch items but not in the Dutch passages. The

main effect of Language on the number of characters in the items ($F(1,14) = 77.80$, $p < .001$) and the passages ($F(1,14) = 9.86$, $p < .01$) was also found to be significant. More characters were thus used in the Dutch passages and items than in the English passages and items, respectively. The main effect of Type of text was not found to be significant in any of the analyses, which shows the number of words and characters in the passages and items do not differ for the literary versus informative texts. A significant interaction between Language and Type of text was also not detected ($p > .05$), which shows the differences according to language to hold across the different types of text.

A two (Type of text: literary vs. informative) by two (Language: Dutch vs. English) analysis of variance was next performed on the mean number of characters per word (mean word length) for the text *passages*. The main effect of Language on mean word length for the passages was significant ($F(1,14) = 68.32$, $p < .001$). The average word length for the Dutch passages was significantly higher than for the English passages. The main effect of Type of text was also significant ($F(1,14) = 33.60$, $p < .001$), which suggests that the mean word length for informative texts was significantly higher than for literary texts. However, a significant interaction between Language and Type of text was also found ($F(1, 14) = 9.66$, $p < .01$). That is, the difference between the Dutch and English versions of the test was significantly larger for the informative texts than for the literary texts with the mean word length for the Dutch informative texts proving highest, followed by the English informative texts, the Dutch literary texts, and then the English literary texts. A Multiple Comparison analysis (Bonferroni corrected) showed all of the differences to be significant at the .01 level, with the exception of the difference between the English informative texts and the Dutch literary texts and the difference between the Dutch literary texts and the English literary texts, which were both non-significant ($p > .05$).

When the mean word length for the *items* from the different versions of the RLT was analyzed, the main effect of Language was non-significant ($p > .05$), but the main effect of Type of text was significant ($F(1,14) = 15.37$, $p < .01$). The informative items contained longer words on average than the literary items. The interaction between Language and Type of text was not significant ($p > .05$).

The mean number of words per sentence (mean sentence length) for the different text *passages* was analyzed next. One passage was not included in the analyses because the passage was a leaflet that contained considerable information in tables and relatively few sentences. A two (Type of text: literary vs. informative) by two (Language: Dutch vs. English) analysis of variance was thus performed on the number of words per sentence for the 15 remaining passages. Neither the main effect of Language ($F(1,13) = 1.30$, $p > .05$) nor the main effect of Type of text ($F < 1$) was significant. The interaction effect between Language and Type of

text was also not significant ($F < 1.0$). The mean sentence length for the items is not presented because each item consisted of only one or two sentences, which is not sufficient to create a mean. The correlations between the number of words used in the Dutch and English versions of the text *passages* and *items* were both highly significant (Pearson's $r = .87$, $p < .001$ and $r = .89$, $p < .001$ respectively). The correlations between the number of characters in the Dutch and English versions of the text *passages* and *items* were also highly significant ($r = .86$, $p < .001$ and $r = .96$, $p < .001$ respectively). When a greater number of words and characters was used in English, thus, a greater number of words and characters was also used in Dutch. The mean word length for the Dutch *passages* also increased linearly with the mean word length for the English passages ($r = .83$, $p < .001$), but no significant correlation was found between the mean word length for the Dutch *items* and the mean word length for the English items ($r = .39$, $p > .05$). Finally, the correlation between the mean sentence length for the Dutch *passages* and the English passages was highly significant ($r = .84$, $p < .001$).

In sum, the Dutch version of the test was found to have more words, more characters, and longer words on average than the English version. The mean word length was also generally longer for the informative texts than for the literary texts, and the difference between the English and Dutch versions of the test was larger for the informative texts than for the literary texts. The correlations between the Dutch and English versions of the test were all significant for the number of words, number of characters, mean word length, and mean sentence length. The only exception to this pattern was the mean word length for the test items: A longer mean word length for the English items was not accompanied by a longer mean word length for the Dutch items.

Complexity

In order to compare the complexity of the Dutch and English versions of the RLT, a panel of six language experts was asked to evaluate the complexity of the test with respect to content, sentence structure, and text structure. More specifically, the experts were asked to compare the complexity of the text passages to a reference passage using a scale that ranged from 1 (= much less complex than the reference text) to 5 (= much more complex than the reference text). The experts were also asked to evaluate the complexity of the test items on a scale from 1 (not complex) to 5 (very complex). The means and standard deviations for the different measures of complexity are presented in Table 3.

Table 3. Means and standard deviations for different measures of complexity for Dutch and English versions of the PIRLS RLT

Language	Content	Sentence Structure	Text Structure	Items
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
Dutch	2.31 (.95)	2.56 (1.03)	2.33 (.73)	2.19 (.78)
English	2.69 (.69)	2.56 (.64)	2.73 (.59)	2.58 (.44)

The complexity of the Dutch version of the RLT did not differ from the complexity of the English version with respect to content ($t(86) = -1.25$, $p > .05$), sentence structure ($t(86) = -1.20$, $p > .05$), text structure ($t(86) = 0.30$, $p > .05$), or test items ($t(86) = -1.06$, $p > .05$). The content complexity of the Dutch version of the test correlated significantly with the content complexity of the English version ($r = .58$, $p < .001$). However, the complexity of the sentence and text structures for the Dutch version of the test did not correlate significantly with the complexity of the sentence and text structures for the English version of the test ($r = .23$, $p > .05$ and $r = .31$, $p > .05$ respectively).

In Table 4, the intercorrelations between the various measures of complexity for the Dutch version of the RLT (upper right corner) and the intercorrelations between the various measures for the English version of the RLT (lower left corner) are presented. As can be seen, all of the complexity measures correlated significantly with each other for both the Dutch and English version of the test.

Table 4. Intercorrelations between various measures of complexity for Dutch version (upper right corner) and English version (lower left corner) of the PIRLS RLT

	Content	Sentence structure	Text structure	Items
Content	–	.71**	.79**	.53*
Sentence structure	.47**	–	.64**	.59**
Text structure	.72**	.68**	–	.58**
Items	.39**	.42**	.50**	–

** significant at the 0.01 level; * significant at the 0.05 level (2-tailed)

Conclusions

Despite considerable efforts to keep the differences between the Dutch and English versions of the PIRLS RLT to a minimum, the translation of the English version of the test into Dutch had some major implications for the linguistic characteristics of the test. While the translators succeeded in creating Dutch passages of text with an equivalent number of words and mean sentence length when compared to the English passages, the children presented the Dutch version of the test nevertheless had to read significantly more characters and longer words than the children

presented the English version of the test. This was found to be particularly the case for the informative passages.

Rather surprisingly, the translation of the test items led to a larger number of words in the Dutch items than in the English items but not to the use of longer words. Nevertheless, the mean word length for the Dutch items did not correlate with the mean word length for the English items. That is, the mean length of the words in the Dutch items did not parallel the mean length of the words in the English items.

In sum, the Dutch children had to read more characters than the English children. In the text *passages*, the Dutch children had to read *longer* words than the English children; in the test *items*, they had to read a *greater number* of words. This finding may have both positive and negative consequences for the performance of the Dutch children. We found the use of a greater number of words and longer words in the Dutch version of the RLT to *not* be associated with more complex sentence and text structures, which means that the Dutch children were clearly not placed at a disadvantage by having to read more complex sentences or texts. The mean complexity of the sentence structures for the Dutch and English versions of the RLT were equal while the mean complexity of the text structure for the Dutch version of the test was a bit lower than for the English version but nonsignificantly so. The Dutch children might still have been at a disadvantage because the reading of a greater number of characters presumably takes more time and cognitive effort, but no evidence was found for this whatsoever. That is, we know from our observations of the testing sessions that the Dutch students had plenty of time to complete the test.

Study 2: Psychometric test properties

The IEA put a tremendous amount of effort into the development of a test to measure the same skill in all participating countries. To also ensure good psychometric properties, the different passages and test items were field tested in 27 countries in September 2000. Based on the results of this field test, the instrument was adapted to create the final versions of the test. In our second study, we therefore examine the consequences of this international adaptation of the instrument for the Dutch version of the test.

Method

Participants

In IEA studies, the target population is referred to as the "international desired target population". For the present PIRLS research, this population consisted of all students enrolled in the upper grade of two adjacent grades containing the largest percentage of 9-year-olds per country at the time of testing (Campbell et al., 2001). In the Netherlands, as in most countries, this was the fourth grade of elementary school (or what is known as *Groep 6* in the Dutch educational system). Students attending schools for special education were not part of this population. The instrument was specifically developed for administration at the end of fourth grade. However, the field test was conducted at the beginning of the school year and we decided to administer it in fifth grade (or what is known as *Group 7* in the Dutch educational system) as we suspected that the test would be too difficult for students just starting fourth grade.

For the PIRLS sample design, it was attempted to test 200 students for each of the 8 booklets from each country (see *Materials*). For this purpose, a field test sample of 70 schools was drawn from the database of all the elementary schools for each country. Given that 100% participation was not possible, an algorithm was used a priori to identify a replacement school for each selected school. In the Netherlands, 47 elementary schools (with 35 schools replacing schools that were initially selected) participated in the field test in the end, which resulted in data on 1470 Dutch students.[2] A total of approximately 48,000 students from almost 1100 schools in 27 countries participated in the field test, which supplied more than 6000 responses for each test booklet.

Materials

The RLT described in Study 1 was used in the field test. The test involved eight booklets with sixteen assessment blocks and 211 test items. Two test items were excluded from the analyses because their scoring rules changed during the scoring process. Each assessment block included either a literary or an informational passage followed by open-ended and multiple-choice questions about the passage.

Procedure

After the field testing of the sixteen assessment blocks, eight assessment blocks were selected for inclusion in the final PIRLS RLT. On the basis of the field test data, the IEA thus omitted eight assessment blocks and 106 test items from the field tested instrument. For the remaining eight blocks, 19 test items were modified, 9 test items were omitted, and 75 test items remained unchanged. The modifications that were made were the same for all participating countries. In order to

study the consequences of the international modifications for the Dutch situation, the Dutch field test data were compared to the international averages of the field test data for the 27 participating countries.

Internal consistency. The Cronbach's alpha reliability coefficients provided information on the internal consistency of the items (i.e., the average inter-item correlation). The international average for each passage was calculated by adding the alphas for the 27 participating countries and dividing this number by 27. The alpha coefficient in the Netherlands was compared with the international average. Further, the alpha coefficient for the selected blocks was compared to the coefficient for the omitted blocks.

Interscorer Reliabilities. Special trained judges assigned a score to the open-ended response item responses (1, 2, or 3 points). In order to assess the reliability of the scoring, about 100 booklets were scored twice for each country. The interscorer reliabilities were the percentages of occasions on which the two scorers assigned exactly the same score to the response of the student. Again, the average inter-scorer reliability in the Netherlands was compared to the international average. Furthermore, the interscorer reliabilities for the unchanged questions were compared to those for the modified and omitted questions.

Item-country interactions. In order to examine the possibility of certain items being more easy or difficult to answer for the students from a particular country (i.e., the possibility of an item-country interaction), the IEA calculated the probability of a hypothetical student with an internationally average level of proficiency correctly responding to an item for each of the test items (based on a Rasch one-parameter IRT model). The average hypothetical probability of a correct response was then compared to the probability of a correct response by a student of average proficiency for each country. A *t*-statistic was next computed by dividing the difference between the country-specific difficulty of the item and the average international difficulty of the item by the standard error for this difference. If a statistically significant negative or positive *t*-statistic was the result ($p < .05$), the item could be considered either unusually easy or unusually difficult for the students from the relevant country, respectively. We calculated a chi-square coefficient to check whether the observed number of unchanged, modified, and omitted items being unusually difficult or unusually easy was equal to or significantly different from the expected number for each cell.

Percentages of correct responding. The mean percentage of correct responses to each item was calculated for the Dutch students and all of the students from the different countries. For the multiple-choice questions, the percentage correct was simply the percentage of the students selecting the correct response for that item. For the open-ended questions worth 1 point, the percentage correct was the percentage of the relevant students scoring 1 point. The percentage correct of

two-points items was calculated by adding up the percentage of students scoring two points, and the percentage of students scoring one point divided by two. The percentage correct of three-points items was calculated by adding up the percentage of the students scoring three points, the percentage of the students scoring two points divided by three and multiplied by two, and the percentage of students scoring one point divided by three.

An analysis of variance was next performed to examine the influence of item omission or modification on the correctness of the responding of the students in the Netherlands and internationally. Finally, the influence of test item type was examined. If differences between countries would be due to the scorers in one country being more generous than in another country, the correctness of the students' responding may vary only for the open-ended items. For this reason, another analysis of variance was performed to examine the influence of test item type (i.e., open-ended or multiple-choice) on the correctness of the students responding in the Netherlands and internationally.

Results

Internal consistency

One of the criteria used to select eight of the sixteen passages for inclusion in the final RLT instrument was the Cronbach's alpha for the responses of the students (i.e., the internal consistency of the test items accompanying the passages). In Table 5, the average alpha coefficients for the sixteen passages as calculated by the IEA for the students from all of the participating countries and the Netherlands in particular are presented. As can be seen, the mean coefficient for the sixteen passages internationally was 0.71 while the mean coefficient for the Netherlands was only 0.61. The mean international coefficient was thus significantly higher than the mean Dutch coefficient ($p < .001$). As might be expected, the mean coefficient for those passages that were kept was higher than the mean coefficient for those passages that were omitted for both the Dutch and international samples. However, analyses of variance showed the difference to be significant for only the international sample ($F(1,14) = 8.24$, $p < .05$) and not the Dutch sample ($F < 1$).

Table 5. Mean Reliability Coefficients for Those Blocks Selected for Inclusion in the Final Version of the RLT, Those Blocks Omitted, and All Blocks

	Netherlands	International
Selected blocks	.63	.73
Omitted blocks	.61	.69
All blocks	.61	.71

Inter-scorer reliabilities

Another criterion used to select passages for inclusion in the final RLT was the degree of inter-scorer agreement for the open-ended questions (i.e., a subsample of the responses to the open-ended questions scored independently by two judges). In Table 6, the average inter-scorer reliability for the 100 open-ended questions can be seen to be almost 85% for the Dutch sample and almost 89% for the international sample. The inter-scorer reliabilities for the Dutch and international samples correlated positively ($r = .78$, $p < .001$). For both samples, the inter-scorer reliability proved highest for the unchanged items. However, analyses of variance showed the differences between the three groups of items to not be significant for either the Dutch sample ($F < 1$) or the international sample ($F(2, 97) = 1.48$, $p > .05$). In other words, the reliability of the scoring for those items that were kept, modified, or omitted did not vary significantly for either the international sample or the Dutch sample.

Table 6. Inter-scorer Reliabilities (and Standard Deviations) for Open-Ended Questions that Remained Unchanged, Were Modified, or Omitted for the Final Version of the RLT and All of the Open-Ended Questions

	Netherlands	International
Unchanged questions (n = 35)	85.7 (11.3)	90.0 (4.1)
Modified questions (n = 14)	84.5 (11.6)	88.0 (4.6)
Omitted questions (n = 51)	84.3 (13.6)	88.3 (5.7)
All open-ended questions (N = 100)	84.8 (12.5)	88.9 (5.1)

Item-country Interactions

As can bee seen from Table 7, 18 of the 209 analyzed items were unusually easy for the Dutch children; 34 items were unusually difficult; and 157 were more or less as expected (i.e., normal) (see *Procedure* for explanation of this classification). It can be further seen that 75 items remained unchanged; 19 items were modified; and 115 items were omitted for the final version of the RLT. Based on these totals, the numbers of expected items for each cell could be calculated. The number of observed items differed significantly from the number of expected items in each cell ($\chi^2_4 = 10.86$, $p < .05$). From the table, it can be seen that for those items that were relatively easy for the Dutch students, fewer items remained unchanged and more items were omitted for the final version of the RLT than might have been expected. For those items that were relatively difficult for the Dutch students, fewer items remained unchanged and more items were modified or simply omitted than might have been expected. For those items that were not particularly easy or difficult for the Dutch students (i.e., normal), more items remained unchanged than might

have been expected. In other words, those items that were not particularly easy or difficult for the Dutch students also went largely unchanged for the final version of the RLT while those items that were relatively easier or more difficult for the Dutch students were modified or omitted for the final version of the RLT more often than expected. The latter finding suggests that the difficulty of the final Dutch version of the RLT may have changed — but not necessarily in the same direction as the final versions of the test for other countries.

Table 7. Item-country Interactions in the Netherlands: Observed versus Expected Numbers of Items

Items	Unchanged	Modified	Omitted	Total
Easy	3 vs. 7	2 vs. 2	13 vs. 10	18
Difficult	7 vs. 12	6 vs. 3	21 vs. 19	34
Normal	65 vs. 56	11 vs. 14	81 vs. 86	157
Total	75	19	115	209

(Expected numbers may not sum to totals due to rounding.)

Percentages of correct responding

In the Netherlands, an average of 67% of the test items were responded to correctly; internationally, an average of 52% of the items were responded to correctly. An overview of the Dutch and international mean scores for the unchanged items, modified items, omitted items, multiple-choice questions, and open-ended questions is presented in Table 8.

A two (Country: Netherlands vs. International) by three (Decision: unchanged, modified, or omitted) analysis of variance was next performed on the percentages of correct responding. Country was treated as a between-subjects variable and Decision was treated as a within-subjects variable. The main effect of Country was highly significant ($F(2, 206) = 299.61$, $p < .001$). The percentage of correct responding in the Netherlands was significantly higher than the percentage of correct responding internationally. The main effect of Decision was also significant ($F(2, 206) = 7.05$, $p \leq .001$). A higher percentage of correct responding was found for the unchanged items than for the omitted or modified items with the difference between the latter not being statistically significant. The Dutch means were consistently higher than the international means, and no significant interaction between Country and Decision was thus found ($F(2, 206) = 2.02$, $p > .05$).

An additional two (Country: Netherlands vs. international) by two (Type of test item: multiple-choice vs. open-ended) analysis of variance was performed on the percentages of correct responding. Country was treated as a between-subjects variable and Type of test item was treated as a within-subjects variable. In Table 8,

the percentages of correct responding for the two types of test items are presented. The main effect of Country was again significant ($F(1, 207) = 506.71$, $p < .001$). That is, the performance of the students in the Netherlands was consistently higher than the performance of the students internationally. The main effect of Type of test item was also significant ($F(1, 207) = 33.29$, $p < .001$). In both the Netherlands and internationally, the performance of the students on the multiple-choice questions was significantly higher than their performance on the open-ended questions. An interaction between Country and Type of test item was not found ($F(1, 207) = 1.27$, $p > .05$).

Table 8. Mean Percentages of Correct Responding (Standard Deviations) in the Netherlands and Internationally

	Netherlands	International
Unchanged items	74.4 (18.6)	57.6 (16.4)
Modified items	57.8 (25.7)	43.3 (16.5)
Omitted items	63.9 (23.6)	50.0 (19.6)
Multiple-choice questions	74.7 (19.5)	59.0 (16.8)
Open-ended questions	58.9 (23.2)	44.7 (17.9)
Total	67.1 (22.7)	52.1 (18.7)

Conclusions

The decision to omit or modify certain items for the final version of the RLT does not appear to have major consequences for the Dutch situation. Internationally, the internal consistency of the blocks selected for inclusion in the final version of the test was higher than the internal consistency of those blocks not selected for inclusion, which means that the selection of eight blocks for inclusion in the final version of the test may actually *increase* the internal consistency of the instrument. Field testing showed the internal consistency of the test items to be rather low internationally and even lower for the Netherlands, which may be explained by the greater number of cases included in the analyses internationally.

The inter-scorer reliabilities for the unchanged, modified, and omitted items did not differ significantly from each other either internationally or in the Netherlands and were very satisfactory. The decision to modify or omit items did not improve the inter-scorer reliability of the test.

The number of items later modified or omitted on the basis of the field results was found to be higher than expected for those items that were found to be more difficult or easier for the students in the Netherlands relative to the international sample. The number of items that remained unchanged for the final version of the

RLT was found to be higher than expected for those items that were not particularly easy or particularly difficult for the students in the Netherlands relative to the international sample. The decision to modify or omit items thus affected those items that did not have good psychometric properties in the Dutch version of the RLT to start with, which means that the contribution of the modifications and omissions is most likely to be positive for the final Dutch version of the test.

Lastly, those items that were later modified or omitted were also the items with the lowest percentages of correct responding in both the Netherlands and internationally. Keep in mind that the percentage of correct responses is not related to the item-country interaction (i.e., unusually easy items in the Netherlands can still have a low percentage correct). The interaction between country and decision to modify, omit, or leave items unchanged was not significant, which means that the difference between the percentage of correct responses for the unchanged items versus the modified and omitted items was the same for the Netherlands as for the international sample. In other words, the modification or omission of items on the basis of the percentage of correct responses will not place the Dutch students at a particular advantage or disadvantage relative to the international sample.

To rule out the possibility of a more lenient scoring process affecting the percentage of correct responses (i.e., the possibility of the Dutch scorers being too generous), the effect of item type on the percentage of correct responses was examined. Recall that the responses to the open-ended items were scored by a team of expert judges following a strict scoring procedure while the correct responses for the multiple-choice items were determined prior to assessment by the PIRLS committee. Our analyses showed the multiple-choice questions to be consistently easier to answer than the open-ended questions irrelevant of country. The fact that the open-ended items were consistently more difficult across countries indicates no bias in the scoring of the responses to the questions in the Netherlands.

General Discussion

Our study suggests that the PIRLS procedure to determine the final (international) version of an instrument to assess the reading literacy of nine- and ten-year olds was stable and will not have negative consequences for the quality of the final Dutch version of the test. The results of the first study reported on here show the Dutch translation of the English version of the RLT to have some inevitable linguistic implications. While the translated Dutch version of the instrument had more characters than the English version, this did not lead to greater sentence or text complexity or to less complex passage content. It should be noted, however, that only six experts evaluated the complexity of the instrument, which meant

only three evaluations per version of a block because the experts judged only half (i.e., eight) of the blocks in English and half (i.e., eight) of the blocks in Dutch. The number of cases per test block may thus have been too small to detect statistically significant differences with regard to test complexity, and future research should certainly address this question.

Differences between languages clearly exist and should not be ignored, particularly in studies of reading literacy. That is, the possibility of linguistic bias in international studies of reading literacy is an important research topic. Differences in the linguistic characteristics of tests can obviously influence student performance and, while the IEA successfully developed a test instrument with largely the same linguistic characteristics for English and Dutch, we cannot say much about the languages other than Dutch. That is, a total of 27 countries participated in the field research, which required almost 30 translations of the test instrument. No empirical conclusions can be drawn about the linguistic similarities and differences between the other versions of the test. Whereas Dutch and English are members of the same Indo-European family of languages and even the same West-Germanic branch of this family, the PIRLS instrument was also translated into languages from the Balto-Slavic branch (e.g., Slovene, Macedonian, and Bulgarian), the Italic branch (e.g., French and Italian), and even the Uralic family (e.g., Estonian and Hungarian) and the Sino-Tibetan family (e.g., Chinese). Languages from the same family often have considerable affinities as they typically stem from the same mother language, which means that those languages resembling the language of the original test instrument may be particularly beneficial — or, for that matter, detrimental — for students. We simply do not know. It would be very interesting to address this question in future research. Indo-European languages play a leading role in today's world and, historically, European nations have implemented their languages in those places where they have come into power. And the same holds for achievement testing and evaluations of literacy, in particular; the possibility of a Western cultural and linguistic bias cannot be ruled out. In contrast to most Indo-European languages, Uralic written languages are based on the phonology of the oral language, which could make the RLT relatively easier for students. In contrast, written Chinese does not form words on the basis of characters; rather, each word has its own symbol (i.e., ideogram) (Vromans, 1988), which could make the RLT relatively more difficult for students. In other words, the linguistic characteristics of the translated versions of the RLT should be empirically examined for different language families and branches of these families in future research.

The results of the second study reported on here suggest that the decisions to modify or omit certain items on the basis of the field test results and to establish the final (international) version of the RLT did not have particularly negative or positive consequences for the psychometric quality of the Dutch version of the

test nor for the average achievement in the Netherlands. The internal consistency of the international English version of the test increased after omission of eight of the sixteen assessment blocks but remained relatively low. In the present study, we only analyzed the data on the average internal consistency of the initial sixteen assessment blocks — as provided by the IEA — and did not analyze the internal consistency of the final eight assessment blocks. The inter-scorer reliability did not change significantly, but was already acceptable. Those items that were modified or omitted for the final version of the test were mostly items that had proved unusually easy or difficult for the students in the Netherlands when compared to the international sample. While it would have been even better for all of the particularly easy or difficult items to be modified or omitted, only ten of the items that were of particular ease or difficulty for the Dutch students remained in the final version of the instrument. The most comforting result is that the decision to modify or omit certain items for the final version of the RLT does not appear to place the Dutch students at a particular advantage or disadvantage with regard to their achievement. Also, the scoring of the open-ended items was found to be clearly unbiased for the Netherlands.

The PIRLS RLT was developed to assess and compare the reading literacy of students around the world. The results of the present validation study have some important implications for future international research on educational achievement. Cultural, linguistic, procedural, and psychometric biases will always threaten the validity of international comparisons. Researchers must therefore put a considerable amount of effort into the prevention of all sorts of biases. To overcome cultural biases, for example, it is important that all of the participating countries be involved in the development of the test instrument from the very earliest stages. To overcome linguistic biases, certain translation standards may be necessary such as having the countries translate a standard international version of the instrument into the target language with as few changes of meaning, difficulty, or layout as possible. Linguistic and psychometric analyses such as those performed in the present study can contribute to the development of an even more valid and reliable instrument. Back-translations might also help identify any further inconsistencies. However, such an undertaking can be very time consuming and expensive. A detailed manual for test administration can help prevent procedural biases. Test administrators should be made aware of the purpose of the test of study, which is the objective assessment of educational achievement to improve education policy. It should also be made very clear that being ranked unjustifiably higher within the international comparison does not help a country. Test administrators should be made aware of the purpose of the test or study, that is, the objective assessment of educational achievement to improve education policy. It should be very clear that a country is not helped by artificially being ordered higher in the international

comparison. Test administrators should recognize the possible impact of any interference during the test sessions and deviations from the standard administration procedures for the comparability of the test results. The test results should also never be used to evaluate school performance as the teachers and school leaders may find it difficult to resist helping the students then. When manual scoring of the student responses is necessary, moreover, the judges must be particularly conscientious, attentive to detail, knowledgeable of the subject area, and willing to apply the scoring guidelines as instructed — even when they might disagree. Finally, psychometric biases can be avoided with the provision of straightforward guidelines for the processing of data and professional software for the keying, checking, and verification of outcomes. Last but not least, thorough pilot testing should be undertaken to demonstrate the reliability and validity of an instrument. A large amount of material should be piloted in order to allow selection on the basis of statistical outcomes (i.e., that material with the most acceptable psychometric properties). Item Response Theories can also help identify those responses that cannot be explained on the basis of the skill assumed to underlie test performance, which was reading literacy within the context of the present study. In short, international comparative study requires the development of a test instrument with good measurement properties for translation into many languages, which takes tremendous effort and expertise.

Acknowledgements

We thank Lee Ann Weeks for her review of our article. We would also like to thank Akke de Blauw, Martine Gijsel, Marjolein Gompel, Eliane Segers, Hanneke Wentink, and Regine van 't Zandt for their participation in the expert panel to evaluate the complexity of the PIRLS Field Test Instrument. Finally, we thank the staffs and fifth-grade students from the elementary schools for their participation in the PIRLS field test in November 2000.

Notes

1. A matrix sampling technique was used to distribute the text passages across the PIRLS students as it was not possible to administer the entire RLT to each child. For the field test, however, this matrix sampling technique was not adhered to because the purpose of the field test was not to compare the reading literacy of the students across countries but to validate the different versions of the RLT.

2. Many of the Dutch schools were not willing to participate in the present study because they had already participated in other (research) projects. Many schools in the Netherlands are also understaffed, which means that the teachers have heavy workloads. In the Netherlands, the

following arrangements were thus made to encourage as many schools as possible to participate in the PIRLS research. Personnel from the Dutch national center conducted the test sessions in the Dutch schools, which deviated from the international procedures that stipulated that the schools themselves should conduct the test sessions. A return envelope was also supplied with the parent questionnaires to allow them to send the questionnaires directly to the national center. Finally, a publication regarding early literacy standards was given to the schools to highlight the importance of the research topic and encourage their participation.

References

Anderson, S., Auquier, A., Hauck, W. W., Oakes, D., Vandaele, W. & Weisberg, H. I. (1980). *Statistical methods for comparative studies. Techniques for bias reduction.* New York, NY: John Wiley & Sons.

Barr, R., Kamil, M. L., Mosenthal, P. & Pearson, P. D. (1990). *Handbook of reading research* (Vol. 2). New York, NY: Longman.

Berk, R. A. (1982). *Handbook of methods for detecting test bias.* Baltimore, MD: The Johns Hopkins University Press.

Bonnet, G., Braxmeyer, N., Horner, S., Lappalainen, H., Levasseur, J., Nardi, E., Remond, M., Vrignaud, P. & White, J. (2001, April). *The use of national reading tests for international comparisons: Ways of Overcoming cultural bias.* Account of a Socrates project, commissioned by the European network of policy makers for the evaluation of education system and co-financed by the European commission. England, Finland, France, Italy.

Camilli, G. & Shepard, L. A. (1994). *Methods for identifying biased tests items.* Thousand Oaks, CA: Sage Publications.

Campbell, J. R., Kelly, D. L., Mullis, I. V. S., Martin, M.O. & Sainsbury, M. (2001). *Framework and specifications for PIRLS assessment 2001.* (2nd ed.). Chestnut Hill, MA: PIRLS International Study Center, Boston College.

Coenen, M., & Vallen, T. (1991). Itembias in de eindtoets basisonderwijs [Item bias in the primary-education final test]. *Pedagogische Studiën, 68,* 15–26.

Extra, G., & Verhoeven, L. (1985). Bias in intelligentie-onderzoek bij allochtone kinderen [Bias in intelligence research of ethnic-minority children]. *Pedagogische Studiën, 62,* 392–395.

Holland, P. W. & Wainer, H. (eds.) (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

IEA (2001). *WinDEM: Software for Data Entry and Verification.* Hamburg: IEA Data Processing Center.

Jong, M. de & Vallen, T. (1989). Linguïstische en culturele bronnen van itembias in de Eindtoets Basisonderwijs voor leerlingen uit etnische minderheidsgroepen [Linguistic and cultural sources of item bias in the final primary-education test for students from ethnic minorities]. *Pedagogische Studiën, 66,* 390–402.

Osterlind, J. O. (1983) *Test item bias. Series: Quantitative Applications in the Social Sciences.* Beverly Hills, CA: SAGE Publications, Inc.

PIRLS (1999). *School Sampling Manual — Version 2.* Prepared by Pierre Foy & Marc Joncas, Statistics Canada. Chestnut Hill, MA: Boston College.

PIRLS (2000). *Survey Operations Manual — Field Test.* Prepared by the International Study Center. Chestnut Hill, MA: PIRLS International Study Center, Boston College.

Sireci, S. G. (1997). Problems and issues in linking assessments across languages. *Educational Measurement: Issues and Practice, 16,* 12–19.

Spilich, G., Vesonder, G., Chiesi, H. & Voss, J. (1979). Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior, 18,* 275–290.

Uiterwijk. H. (1994). *De bruikbaarheid van de Eindtoets Basisonderwijs voor allochtone leerlingen* [The usefulness of the primary-education final test for ethnic-minority students]. Arnhem, the Netherlands: Instituut voor Toetsontwikkeling (Cito).

Vromans, J. (1988). *Taal. Het grote avontuur.* [Language. The big adventure]. (H. J. Storig, Trans.). Utrecht, the Netherlands: Het Spectrum. (Original work published 1987)

Authors' addresses:

Mieke van Diepen
Expertisecentrum Nederlands [National Center for Language Education]
P.O. Box 9104
6500 HE Nijmegen
The Netherlands
Tel: +31 24 36 15 624
Fax: +31 24 36 15 644

E-mail: m.vandiepen@pwo.ru.nl

Mieke van Diepen, Ludo Verhoeven, and Cor Aarnoutse,
National Center for Language Education [Expertisecentrum Nederlands]
Radboud University Nijmegen

Anna M.T. Bosman,
Department of Special Education
Radboud University Nijmegen.