MEASUREMENT: INTERDISCIPLINARY RESEARCH AND PERSPECTIVES
2016, VOL. 14, NO. 1, 1–26
http://dx.doi.org/10.1080/15366367.2015.1112711

Routledge
Taylor & Francis Group

FOCUS ARTICLE

Linking Large-Scale Reading Assessments: Measuring International Trends Over 40 Years

Rolf Strietholt[a,b] and Monica Rosén[c]

[a]Institute for School Development Research, Techniche Universität Dortmund; [b]Center for Educational Measurement, University of Oslo, Oslo, Norway; [c]Department of Education and Special Education, University of Gothenburg

ABSTRACT

Since the start of the new millennium, international comparative large-scale studies have become one of the most well-known areas in the field of education. However, the International Association for the Evaluation of Educational Achievement (IEA) has already been conducting international comparative studies for about half a century. The present study aims to demonstrate how to link recent and older studies onto the same scale in order to study long-term trends within and across countries. It discusses the comparability of the assessment material in the Progress in International Reading Literacy Study (PIRLS) and previous IEA studies on reading at the end of primary school. Thereafter, we use a concurrent calibration of all item parameters to link the studies onto a common IRT scale extending from 1970 until the present.

KEYWORDS

item response theory; linking study; PIRLS

Background

Because many features of educational systems and educational policies vary only across countries, international comparative studies provide a unique approach to analyzing the impact of specific educational policies on educational outcomes (Hanushek & Wößmann, 2011). However, this idea that one could "use the world as an educational laboratory" was already formulated about half a century ago. Since then, organizations such as the IEA (International Association for the Evaluation of Educational Achievement) and the OECD (Organisation for Economic Co-operation and Development) have conducted about 50 international comparative studies and collected a rich body of data on student achievement in different subjects and age groups (Wendt, Bos, & Goy, 2011). The results of these studies, however, have been used mostly for cross-sectional comparisons but not for studies of trends on a country level. The main purpose of the present article is to demonstrate how to calibrate the achievement scales from past and present studies on a common IRT scale. This delivers an empirical basis for investigating the long-term effects of educational policy and policy-related issues on educational outcomes. Such data are particularly useful in the field of education, because many reforms (e.g., of teacher training) need to have been implemented fairly extensively and for a certain time before they would reveal an impact on educational outcome.

Related research

To understand why international trend analyses are rare, it seems worthwhile to recapture the background and development of the international studies. Currently, the largest recent international assessments are the IEA studies PIRLS and TIMSS and the OECD's PISA. Because these studies are designed to measure change, they apply, for example, a common assessment framework to ensure

CONTACT Rolf Strietholt Rolf.strietholt@tu-dortmund.de Institute for School Development Research, Technische Universität Dortmund, Vogelpothsweg 78, Dortmund 44227, Germany.

that the study cycles have identical target samples and similar tests in order to make results comparable (Gonzales & Rutkowski, 2010; Gustafsson, 2008; Strietholt, Rosén, & Bos, 2013). However, studying trends was not a priority in the early days of international comparative education. This restricts the comparability of the results from the recent and older studies. Looking at reading at the end of primary school, for instance, one can see that trend studies with PIRLS data are limited to the past 10 years because the study was first implemented in 2001. The present study aims to extend this period by linking PIRLS with earlier IEA studies spanning a period from 1970 until now.

The development of international studies and the consequences for the comparability of the assessments are best described with concrete examples. For the sake of simplicity, we focus on reading at the end of primary school. Up to now, the IEA has implemented three major reading assessment programs in this field. The first was the Reading Comprehension Study (RCS) from 1970; the second, the Reading Literacy Study (RLS) from 1991 and its repetition in 2001; and the most recent, the Progress in International Reading Literacy Study (PIRLS), carried out every 5 years since 2001 (Elley, 1992; Martin, Mullis, Gonzalez, & Kennedy, 2003; Mullis, Martin, Foy, & Drucker, 2012; Mullis, Martin, Gonzales, & Foy, 2003; Mullis, Martin, Kennedy, & Foy, 2007; Thorndike, 1973). Table 1 shows that 9–49 countries participated in the respective studies and that they cover a time period of more than 40 years.

Studying trends with data from recent and older studies

To explore the possibilities and limitations of investigating long-term trends across studies, it is worth taking a closer look at international study reports. The RCS-1970 report covers raw scores––average numbers of correct responses––that range between 3.7 points in Iran and 21.5 points in Sweden (Thorndike, 1973). The scores in the RLS-1991/2001 report are IRT scores with an international mean of 500 and a standard deviation of 100. The PIRLS report also covers IRT scores with a 500/100 scale (Martin et al., 2003, 2012; Mullis et al., 2007), but the scale differs from the RLS-1991/2001 scale because the standardizations and transformations onto the 500/100 scales are based on different countries. It is quite obvious that we cannot use the scores from the international reports to study long-term trends because the reading scale does not have the same metric.

The current study

The purpose of this study is to demonstrate how to link older and recent international comparative studies on reading achievement at the end of primary school. We use item response theory (IRT) to link the IEA studies RCS, RLS, and PIRLS onto a common metric for all studies from 1970 until now in order to investigate long-term trends in student achievement on the country level. The basic design of our study was as follows: (a) to investigate similarities and differences across the different tests, (b) to identify overlaps in the test material, (c) to calibrate all item parameters by means of IRT models, and (d) to adjust the scores for differences in the sample compositions. For the sake of simplicity, we have chosen to focus on the 4 countries Hungary, Italy, Sweden, and the United States because these are the only countries that participated in all studies.

Table 1. IEA Studies on Reading at the End of Primary School.

	Test	Study	Years	Countries	Target sample
1	RCS	Reading Comprehension Study	1970	14	10-year-olds
2	RLS	Reading Literacy Study	1991	26	Grade with the most 9-year-olds
			2001	9	
3	PIRLS	Progress in International Reading Literacy Study	2001	35	Grade 4
			2006	40	
			2011	49	

Comparing RCS, RLS, and PIRLS

Meaningful comparisons across different assessments require comparable data. Kolen and Brennan (2004) proposed four criteria to evaluate the degree of similarity between tests: inferences, populations, constructs, and measurement characteristics/conditions. We use this scheme to explore similarities and differences in the assessment material of the studies before discussing the consequences for linking the tests.

To what extent are the assessments used to draw similar inferences?

The purposes of assessments have important consequences for their comparability across studies; for example, Feuer, Holland, Green, Bertenthal, and Hemphill (1999) discuss whether it is feasible to link state or commercial tests to the NAEP (National Assessment of Educational Progress) or to TIMSS (Third International Mathematics and Science Study). They are rather skeptical because NAEP and TIMSS have low stakes for test takers. In contrast, when stakes are high, students are usually more motivated to try harder or teachers are more likely to adapt their instruction to fit a specific test. The current study is not influenced by different purposes of the various assessments because the respective assessments were designed to compare the performance of different educational entities.

Are the assessments designed for the same population?

All assessments address students at the end of primary school. Thus, the target population is quite similar across all studies. However, in international student achievement studies, it is not possible to define the target population in such a way that both age and grade are balanced across all countries because of their different school-entry ages. A comparison of the target sample definitions reveals that the study designers employed different approaches. In the RCS-1970, for example, the target sample was defined as "students aged 10:0–10:11 years at the time of testing" (Postlethwaite, 1974, p. 164). This means that in countries with an early school-entry age, the samples were composed of students from higher grades in comparison to countries in which students enter school later. PIRLS sampled "students enrolled in the grade that represents four years of schooling" (Joncas, 2007, p. 36). This increases the international comparability in terms of grade while, at the same time, reducing comparability in terms of age. The designers of RLS even employed a combined age- and grade-based sampling approach: "The grade level in which most students were ages 9:00–9:11 years" (Elley, 1994, p. 239).

How similar are the test conditions?

All studies gave the examinees text passages together with questions based on the passages. Students had to read a passage and then answer the questions while referring back to the passage if necessary. In RCS-1970 and RLS-1991/2001, each student took all text passages. Identical tests were used in both RLS cycles in order to compare 10-year trends. PIRLS used a complex matrix sampling technique to divide the test material into a number of booklets. Each booklet contained two text passages and every passage appeared in multiple booklets. Students worked on one booklet so that information on 2 text passages and the respective items was available for every individual (Martin, Mullis, & Kennedy, 2003). All in all, the PIRLS-2001 tests contained 8 passages; the 2006 test, 10 passages; and the 2011 test, 10 passages. This approach made it possible to extend the assessment material because each student worked on half of the items only. This is advantageous for assessing test validity. Furthermore, the matrix sampling technique was used to link the different assessment cycles by integrating 4 of the text passages from PIRLS 2001 into PIRLS 2006 (Martin, Mullis, & Kennedy, 2007). The assessment material from PIRLS 2011 contains 2 passages that were used in 2001 and 2006 and another 4 passages that were used in PIRLS 2006 only (Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009).

All studies administered the tests in 2 test sessions with a short break in between the sessions. In the RCS-1970, the time allowance was 50 minutes for 45 multiple-choice items; in the RLS-1991/2001, 75 minutes for 66 multiple-choice items. In each PIRLS cycle, the test time allowance per student was 80 minutes. The tests were composed of 98 items in 2001, 126 items in 2006, and 135 items in 2011. Due to the matrix sampling technique, each PIRLS student did not work on all but only about 50 of the items. PIRLS students had more time because the test contained not only multiple-choice but also constructed-response items that require students to formulate their own answers. Such items were worth 1, 2, or 3 points depending on how elaborated the student's answer was. The item formats mark a difference between the tests, because the RCS and RLS tests contained no constructed-response items.

To what extent do the tests assess a similar construct?

According to the test developers, the earlier studies served as a model for the later studies (Campbell, Kelly, Mullis, Martin, & Sainsbury, 2001; Elley, 1994). However, we have to investigate the definitions and operationalizations in more detail to evaluate to what extent the tests assess a similar construct. For this purpose, we shall take a closer look at the respective assessment frameworks and tests.

> The PIRLS-2001 assessment framework defines reading literacy as the ability to understand and use those written language forms required by society and/or valued by the individual. Young readers can construct meaning from a variety of texts. They read to learn, to participate in communities of readers, and for enjoyment. (Campbell et al., 2001, p. 3)

The respective test passages and questions were classified in terms of their purposes for reading and the comprehension processes assessed. PIRLS-2001 distinguished between 2 reasons why young children read: for "literacy experience" and to "acquire and use information." After reading, students have to demonstrate a range of skills and abilities in constructing meaning from the passages. The different questions address 4 hierarchically ordered processes of reading comprehension: (a) "focus on and retrieve explicitly stated information," (b) "make straightforward inferences," (c) "interpret and integrate ideas and information," and (d) "examine and evaluate content, language, and textual elements." PIRLS was designed to measure trends and, therefore, the tests from the 2006 and 2011 cycles employed the same definition of reading literacy. About half of the text passages and items from previous PIRLS cycles were systematically integrated into the assessment material of the subsequent cycles (Mullis, Kennedy, Martin, & Sainsbury, 2006; Mullis et al., 2009). Roughly speaking, the different purposes and processes were equally represented in both PIRLS cycles.

Elley's (1994, p. 5) definition of reading literacy for RLS is similar: "the ability to understand and use those written language forms that are required by society and/or valued by the individual." However, a closer look reveals a crucial difference from the PIRLS test, because the RLS test covers not 2 but 3 text types (Binkley, 1994). The first 2 categories, "narrative" texts and "expository" texts, are basically the same text types as in PIRLS. Narrative texts are continuous texts that aim to tell a story, and students read them for literacy experience. Expository texts describe or explain something, and students read such material to acquire and use information. The third category is the so-called documents, and they are not well represented in PIRLS. Documents refer to noncontinuous tables, charts, graphs, and other documents such as maps or bus schedules. These types of texts also introduce a fifth reading process: "locating some specified information in a structured document" (e.g., choosing the correct bus route). The other 4 reading processes are also well represented in RLS. The RCS test aimed to measure

> [the] ability to determine the meaning of a word or phrase in context, to answer questions that are specifically answered in the passage, to draw inferences from a passage about it contents, and to determine the writer's purpose, intent and point of view. (Thorndike, 1973, p. 56)

As mentioned previously, the assessment frameworks and the technical reports of the RLS as well as PIRLS referred to the early RCS (cf. Binkley, 1994; Campbell et al., 2001). The RCS test covered narrative and expository text types and questions that address the processes of reading comprehension already introduced above.

Degrees of similarity and consequences for linking

Table 2 presents an overview of the distribution of the text types and the reading processes in the different assessments. All in all, we can see a high degree of similarity across the IEA studies but also some differences. In this section, we discuss the consequences of the differences for the comparability of the respective studies. Furthermore, we elaborate on how the present linking study should overcome this limitation.

The most obvious difference in the test contents are the "document" type texts that appear in the RLS-1991/2001 test but in no other study. Gustafsson and Rosén (2006) show that the document-related items introduce a systematic source of variance into the tests. This finding implies that we have to be careful when comparing results across various tests. To overcome possible biases, we suggest disregarding the documents in the present study. In contrast, the proportion of texts for "literacy experience" and to "acquire and use information" is roughly the same in all 3 assessments.

The proportion of the various processes of reading comprehension marks another difference, because these were not represented equally in the assessments. For example, the RCS study contained relatively few simple items that required the students to determine the meaning of a word or phrase in context. Such differences indicate differences in the difficulty levels of the tests because the underlying processes have a hierarchic order (Campbell et al., 2001). A test with a high proportion of items requiring students to retrieve explicitly stated information is easier than a test in which the student has to determine the writer's point of view. We employ item response theory (IRT) to estimate the item and test difficulties so that we can interpret reading literacy on the same scale (e.g., Hambleton, Swaminathan, & Rogers, 1991).

Another criterion is the test conditions—that is, the item formats and test length. The assessment material contained multiple-choice and open (constructed response) questions that are worth up to 3 points. The latter item format was well represented in the 3 PIRLS cycles but not in any previous study. Such differences in the test conditions can influence comparability because students may guess the correct answers in multiple-choice tasks. Here we can also take advantage of using IRT to analyze the data. Mixing items can lead to unequal weightings on total scores. In IRT, however, mixed item formats pose no difficulty because we can model guessing and different numbers of categories (Embretson & Reise, 2000). The test length varied across the studies, but this influences only the standard errors of measurement and not the point estimates. In IRT, this error is not assumed to be the same across all test takers but specific to students with certain abilities and to depend on the number of items and the item properties.

Finally, changes in the target population are another systematic source of variation. Due to different school entry policies, the respective samples were composed of 3rd, 4th, or even mixed grades and the samples also differed in terms of their average age. To understand the consequences of these changes, we have to distinguish between what the tests measure (test contents) and the comparability of the results across countries or over time. Concerning the test contents, we do not think that the test developers were influenced by changes in the target population, because all studies addressed students at the end of primary school. However, changes most certainly influenced the comparability of how well students performed. For example, Hungary tested 3rd graders in RLS-1991/2001 but 4th graders in PIRLS-2001/2006/2011. It is beyond the primary scope of this paper to elaborate on this issue in great detail. However, in another article, we have explored the possibilities of adjusting the observed average achievement scores for differences in terms of age and grade (Strietholt, Rosèn, & Bos, 2013).

Table 2. Degree of Similarity for Different Tests.

Test contents and conditions	RCS	RLS	PIRLS
Text types	3 literary experience 5 acquire and use in-formation	4 literary experience 5 acquire and use in-formation 6 documents	2001: 4 literary experience 4 acquire and use information 2006: 5 literary experience 5 acquire and use information 2011: 5 literary experience 5 acquire and use information
Processes of comprehension	Determine the meaning of a word/phrase in context (8%) Answer questions that are specifically addressed in the passage (33%) Draw inferences from a passage about its contents (46%) Determine the writer's purpose, intent, and point of view (13%)	Match the word of the item to that in the text (17%) Choose/compose an answer that is stated explicitly in the text (24%) Identify the main theme or underlying message of the text (6%) Interpret and integrate ideas and information (2001: 40%; 2006: 37%; 2011: 28%) Generalize from the text about some character or event (18%) –only for the documents;	Focus on and retrieve explicitly stated information (2001: 22%; 2006: 22%; 2011: 24%) Make straightforward inferences (2001: 24%; 2006: 28%; 2011: 34%) Evaluate content, language, and textual elements (2001: 14%; 2006: 14%; 2011: 13%) Search and find some specified information contained in a structured document (35%)
Time allowance	25 + 25 min	35 + 40 min	40 + 40 min
Items[a]	45 MC	66 MC (23 of them are related to the documents)[b]	2001: 98 (46 MC, 44 CR) 2006: 126 (64 MC, 62 CR) 2011: 135 (74 MC, 61 CR)

[a]MC = multiple choice; CR = constructed response.
[b]The RLS test covered 2 CR items, but they have not been scored and are not available in the data sets. *Sources.* Binkley (1994); Campbell et al. (2001), Martin et al. (2007), Martin and Mullis (2012), and Thorndike (1973).

Method

Data

The present study employs data from the 6 IEA studies RCS-1970, RLS-1991/2001, and PIRLS-2001/2006/2011. For the sake of simplicity, our study focuses on data from Hungary, Italy, Sweden, and the United States because these are the only countries that have participated in all previous assessments. All studies employed complex sampling designs in which countries sampled students or entire classes within schools. The data sets cover sampling weights to compensate for unequal sampling probabilities. The current study is based on data from 105,910 students. Although all samples were composed of students at the end of primary school, they differed to a certain degree in terms of age and schooling because the school entry age varies across countries and there were changes in the target sample definitions. In PIRLS 2001, Sweden extended the international study with an additional grade 3. We included this data in our study. The data and documentation files for RLS-1991/2001 and PIRLS-2001/2006/2011 are available online at the IEA Study Data Repository (http://rms.iea-dpc.org/). Data from the older study RCS-1970 have been made available online through the Center for Comparative Analyses of Educational Achievement (COMPEAT; http://www.ips.gu.se/english/Research/research_databases/compeat/). The COMPEAT environment also provides the data for the Swedish extension in RLS-1991 (see below).

Instruments

The reading tests in all studies contained 40 text passages and 344 questions based upon them (see Appendix A). Due to the exclusion of the 6 document texts from RLS as well as the 2 PIRLS Reader-texts in the respective PIRLS cycles, our analyses are based on 28 text passages and 238 test items. We will describe the tests as well as our reasons for excluding some texts in the remaining part of this section.

The RCS-1970 is composed of 8 passages and 45 items to assess reading comprehension. We conducted a factor analysis to examine the dimensionality of the RCS test. The eigenvalue of the first factor is 7.5, which is considerably greater than the second eigenvalue of 1.5 and those for the remaining factors (Appendix B illustrates the distribution of the eigenvalues).

The RLS test is composed of 15 texts (66 items) of which 6 (23 items) were "document" type texts. We disregarded the documents in our linking study because this text type is not well represented in the assessment material of the other studies (see section "Degrees of similarity and consequences for linking"). We conducted a factor analysis using the combined RLS data from 1991 and 2001 for the remaining 9 texts (43 items). Again, the eigenvalue of the first factor (10.2) is considerably greater than those of the remaining factors (the second eigenvalue is 2.3).

The tests in PIRLS contained 8 texts (98 items) in 2001, 10 (126 items) in 2006, and 10 (135 items) in 2011. PIRLS used a matrix sampling technique to distribute the assessment material into booklets. Each text appears in multiple booklets to enable linking among booklets, but the respective texts are *not* paired with one another. One consequence of this approach is that the correlation matrices of the test items have structural missing data by design. For this reason it is not feasible to compute eigenvalues to examine the dimensionality of the PIRLS tests.

The assessment material in each PIRLS cycle covered one "Reader." The Readers were full-color and magazine-style booklets composed of 2 different texts in each PIRLS cycle. We excluded these texts because they appeared only in the Reader but in no other booklet. Thus, there are no overlaps with the other assessment materials. Excluding the Readers reduces the number of texts to 6 (74 items) in 2001, 8(99 items) in 2006, and 8 (102 items) in 2011. We also had to exclude all students who took the Reader because no other texts were administrated to them. Students who took the Reader made up approximately 25% of the PIRLS samples (for further information see Martin & Mullis, 2012; Martin et al., 2003, 2007).

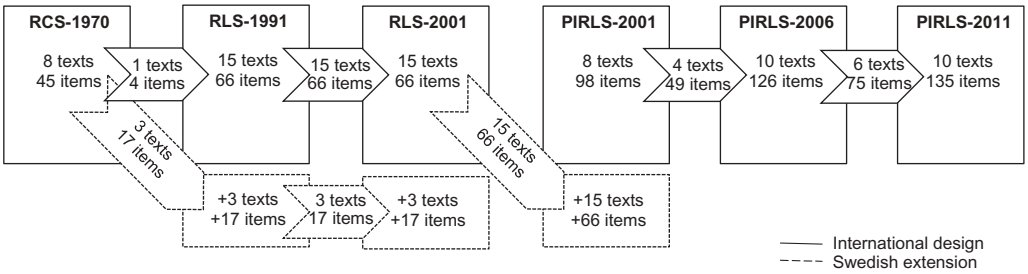


Figure 1. Overlaps across the tests in the international design and in the Swedish extension.

Note. The figure illustrates the number of text passages and the corresponding items and the Swedish extensions to the international design. The arrows indicate the number of texts and items that were re-administered in subsequent studies.

In the next section, we elaborate on the overlaps across the different tests. Figure 1 illustrates that some of the text passages and items from earlier studies were integrated or readministered in subsequent studies. Here we have to distinguish between the "international design" and "Swedish extensions" of the international design. The Swedish national research coordinators added some valuable extensions to the studies from 1991 and 2001. These extensions allow us to link the tests from all assessments. We shall describe the international design of the studies and show limitations for a linking study before elaborating on the extensions.

Overlaps in the assessment material: The international design

Only one passage of the RCS-1970 test ("Marmot," 4 items) was integrated into the assessment material of the RLS-1991/2001 test that contained 15 texts in total. The same test was used in the both RLS cycles. There are no overlaps between PIRLS and the previous IEA studies. PIRLS integrated linking items from previous PIRLS cycles into the assessment material of each cycle. The PIRLS-2001 assessment material contained 8 passages, 4 of which were integrated into the PIRLS-2006 test, which contained a total of 10 passages. The assessment material from 2011 contained 10 texts: 4 newly developed passages and 6 old passages from 2006 of which 2 also appeared in the 2001 test. The overlaps between 2001 and 2006 (= 49 test items), 2006 and 2011 (= 75 items), and 2001, 2006, and 2011 (= 25 items) make it possible to link the 3 tests onto the same scale and compare the results of all 3 PIRLS cycles.

Overlaps in the assessment material: The Swedish extensions

Sweden extended the assessment material in both RLS cycles by giving 3 additional text passages from the RCS-1970 material ("Erneke," "Plant," and "Tailor": 17 items) to the Swedish students (Taube, 1993). Together with the "Marmot" text, the overlap between RCS-1970 and RLS-1991 involves 4 text passages (= 21 items) in Sweden. Furthermore, each PIRLS-2001 student in Sweden took either the first or the second half of the RLS test (Gustafsson & Rosén, 2006).

Age and grade

Teachers or school coordinators recorded information on student age and grade during the process of data collection.

Missing data

It is useful to distinguish between 3 types of missing data in achievement tests: not-administrated, omitted, and not-reached items. In the present study, not-administrated items were treated as missing data in estimating item and student parameters. The omitted items were treated as incorrect responses, because we did not want to reward students for skipping an item. Sometimes, however,

students run out of time and do not complete the items at the end of the test. The not-reached items are typically treated as if they were not administrated when estimating the item parameters. However, they may be treated either as incorrect or as if they were not administrated when student proficiency scores are generated. The first approach punishes slow students. In the present study, we adopted the second alternative because the proportion of students who did not finish the tests varied considerably between the studies: 8% in RCS-1970, 25% in RLS-1991, 31% in RLS-2002, 7% in PIRLS-2001, 5% in PIRLS-2006, and 4% in PIRLS 2011. The observed differences indicate that reading speed is a more significant systematic source of variance in RLS-1991/2001 than in the other studies. Such differences limit the comparability of the tests (see Gustafsson & Rosén, 2006). We avoid this issue by treating the not-reached items at the end of the tests as if they were not-administrated.

In general, the amount of missing data on students' age and grade was very small (< 0.2%).

Linking method

The present study had a common-item nonequivalent group design in which no text passage appeared in all studies but a set of common texts appeared in multiple studies (Kolen & Brennan, 2004, see Figure 1). Item response theory (IRT) models were used to link all tests onto the same metric by means of a concurrent calibration of all item parameters (Kim & Cohen, 1998, 2002). The procedures and models used in the present study are similar to the modeling approaches in modern IEA studies (see, for a detailed description of the procedures, Gonzalez, 2003; Gonzalez, Galia, & Li, 2004). Different IRT models were used for different item types. For multiple-choice items, we used a 3-parameter logistic (3PL) model that gives the probability that a student s with the unobserved reading ability Θ gives the correct answer to item i:

$$P(x_{is} = 1|\Theta_s, a_i, b_i, c_i) = c_i + \frac{1 - c_i}{1 + \exp(-1.7a_i(\Theta_s - b_i))} \tag{1}$$

in which

x_{is} is the response of student s to item i (0 or 1 if correct),
Θ_s is the reading ability of student s,
a_i is the slope/discrimination parameter of item i,
b_i is the location/difficulty parameter of item i, and
c_i is the lower asymptote/guessing parameter for item i.

Another category of test items comprises so-called constructed-response items. Students cannot guess among the distractors but have to formulate and write their answer. Thus, the guessing parameter c was fixed at zero for dichotomous constructed responses resulting in the 2-parameter logistic (2PL) model.

The third type comprised polytomous items that were worth 2 or 3 points but for which the students could also receive partial credit for partially correct responses. We used the generalized partial credit model (Muraki, 1992) for such items. The basic idea is to model the probability for each response category of the polytomous item i:

$$P(x_{is} = l|\Theta_s, a_i, b_i, d_{i,1}, \ldots, d_{i,m_i-1}) = \frac{\exp\left[\sum_{v=0}^{1} 1.7a_i(\Theta_s - b_i + d_{i,v})\right]}{\sum_{g=0}^{m_i-1} \exp\left[\sum_{v=0}^{g} 1.7a_i(\Theta_s - b_i + d_{i,v})\right]} \tag{2}$$

in which

m_i is the number of response categories for item i,
x_{is} is the response of student s to item i (ranging between 0 and m_i-1),
Θ_s is the reading ability of student s,

a_i is the slope/discrimination parameter of item i,
b_i is the location/difficulty parameter of item i, and
$d_{i,l}$ is the category l threshold parameter of item i.

The calibration of the item parameters and the estimation of students' ability are 2 separate steps. First, we calibrated the item parameters in Equations (1) and (2) onto the same IRT scale using data from all studies and countries. The R package Test Analysis Modules (TAM, Kiefer, Robitzsch, & Wu, 2015) was used for the multiple group IRT analyses. The program uses the expectation-maximization algorithm to achieve marginal maximum likelihood estimates of the item parameters. Latent normal distributions of student proficiency are assumed for each country in the respective studies. As RLS-2001 and PIRLS-2001 were administrated in the same year, more than 1 sample is available for each country. In this case we treated both samples as 1 group if a country sampled the same grade in both studies. For example, the United States samples 4th graders in both studies. However, we modeled the samples as 2 separate groups if a country sampled different grades. For instance, Hungary sampled third graders in RLS-2001 and fourth graders in PIRLS-2001. The combined Swedish third grade data from RLS-2001 and PIRLS-2001 functions as a reference group in which the mean is fixed to zero with a variance of 1; means and variances are freely estimated for the other groups.

Then, we used these item parameter estimates to compute students' reading ability scores. To take the estimates' uncertainty into account, we draw 5 plausible values (PVs) from the empirically derived posterior distribution of reading proficiency for each student (von Davier, Gonzalez, & Mislevy, 2009). We repeat all further analyses 5 times, once with each of the PVs, and combine the results using Rubin's (1987) rules. We transformed each PV on a metric with a mean of 500 and a standard deviation of 100 points. We used the transformed scores to compute the mean reading achievement for the respective country × study samples using sampling weights. We employed the jackknife repeated replication technique to estimate the sampling error for the complex samples (Lohr, 2009; Wolter, 2007). We used 2 different versions of the jackknife technique: The more recent studies, RCS-2001 and PIRLS-2001/2006/2011, contained so-called jackknife zones and replicated weights that we used to employ the jackknife 2 (JK2) technique. Because the older studies, RCS-1970 and RLS-1991, did not contain such information, we used the more general jackknife 1 (JK1) procedure in these studies.

Results and discussion

The results of our analysis are summarized in Table 3 that lists the countries' average reading achievement scores from the international reports (*Original_Scores*) along with scores on the new common scale (*IRT_Scores*). The IRT parameters for each item are listed in Appendix C. The overview reveals a considerable variation in the location ($M = -0.793$; $SD = 1.010$; $N = 238$), slope ($M = 1.061$; $SD = 0.361$; $N = 238$), and the guessing parameters ($M = 0.170$; $SD = 0.113$; $N = 163$). The variance in the difficulty (location) parameters indicates that the assessment material covers a broad ability range; such a variance can be observed across all studies. The negative value of the average item difficulty indicates that the test was relatively easy for the test takers in the reference group. The variation in the slope and guessing parameters provides empirical support for our model choice in comparison to more restrictive models. For example, the 1-parameter logistic (1PL) model assumes equal item slopes and no guessing. Obviously enough, the estimated item parameters in the present studies do not support these assumptions.

We further evaluate the model fit by the Q3 statistic, which is the correlation between the residuals of item pairs (Yen, 1984). As we use data from i = 238 items, there are 28,203 possible combinations. However, it is not possible to compute this statistic for all item pairs because not all items were administrated together (structural missing by design). Therefore, the distribution of Q3 statistics refers to the available data of 10,987 item pairs: the mean is −0.031 with a standard deviation of 0.041 ($Q3_{min} = -0.254$, $Q3_{max} = 0.212$). In 6 cases the correlations are beyond the range of −0.2 and 0.2. It seems worthwhile to mention that the highest positive correlations can be observed among items from

Table 3. Results from the Test Linking.

			Grade		Age		Average reading achievement			
							Original_Scores[a]		IRT_Scores	
Study	Country	N^b	*M*	*SD*	*M*	*SD*	*M*	*SE*	*M*	*SE*
RCS 1970	Hungary	4,845	4.29	0.48	10.58	0.28	14.0	n.a.	471	3.8
	Italy	4,466	4.96	0.23	10.59	0.31	19.9	n.a.	551	4.3
	Sweden	1,951	3.53	0.50	10.45	0.31	21.5	n.a.	543	4.0
	USA	5,420	4.66	0.53	10.62	0.38	16.8	n.a.	488	3.5
RLS 1991	Hungary	3,009	3.00	0.00	9.34	0.56	459	4.0	457	3.4
	Italy	2,232	4.00	0.00	9.85	0.46	500	5.4	487	4.3
	Sweden	4,303	3.00	0.00	9.75	0.31	513	4.2	509	3.1
	USA	6,546	4.00	0.00	9.99	0.56	521	3.2	496	3.5
RLS 2001	Hungary	4,707	3.00	0.00	9.69	0.56	475	3.9	471	3.4
	Italy	1,590	4.00	0.00	9.86	0.33	513	4.4	495	3.8
	Sweden	5,361	3.00	0.00	9.79	0.37	498	3.9	495	2.9
	USA	1,826	4.00	0.00	9.96	0.70	511	6.3	486	6.4
PIRLS 2001	Hungary	3,502	4.00	0.00	10.67	0.52	543	2.2	502	2.5
	Italy	2,628	4.00	0.00	9.85	0.37	541	2.4	495	2.9
	Sweden	6,216	4.00	0.00	10.80	0.33	561	2.2	540	2.3
	Sweden	9,385	3.00[d]	0.00	9.81	0.34	520	2.7	486	2.2
	USA	2,837	4.00	0.00	10.22	0.45	542	3.8	497	4.7
PIRLS 2006	Hungary	3,269	4.00	0.00	10.66	0.48	551	3.0	510	3.5
	Italy	2,861	4.00	0.00	9.70	0.35	551	2.9	510	3.4
	Sweden	3,536	4.00	0.00	10.85	0.33	549	2.3	509	2.7
	USA	4,131	4.00	0.00	10.08	0.51	540	3.5	492	4.1
PIRLS 2011	Hungary	4,142	4.00	0.00	10.67	0.50	539	2.9	493	3.1
	Italy	3,322	4.00	0.00	9.73	0.36	541	2.2	492	2.6
	Sweden	3,683	4.00	0.00	10.74	0.34	542	2.1	493	2.8
	USA	10,142	4.00	0.00	10.23	0.44	556	1.5	516	1.8

[a]The scores and standard errors of the overall reading scale are taken from the international study reports (Martin et al., 2003, 2012; Mullis et al., 2007; Thorndike, 1973)
[b]*N* refers to the sample size in the present study.
[c]Standard errors (*SE*) were not available (n.a.) in the RCS-1970 report.
[d]Sweden extended the international study design that targeted Grade 4 with an additional Grade 3 sample.

the same booklet of the RLS test. The highest negative correlations can be observed among multiple choice and constructed response items. The information from the piecewise assessment of model fit may be summarized in an overall goodness-of-fit measure. The standardized root mean square residual (SRMSR) is the square root of the average of the squared residual correlations. Maydeu-Olivares (2013) suggests that values smaller than 0.05 indicate a substantively negligible misfit. In the present study, we observe a SRMSR = 0.039, which indicates an acceptable fit.

Both scales can be used to compare the rank order of the countries. For example, both scales indicate that Hungary was the lowest performing country in 1970 and 1991. Over and above such relative comparisons, the main advantage of the new *IRT_Scores* is their comparability across studies. Linking the tests adjusts for differences in the measurement scales so that we can use the results to evaluate long-term trends on the same ability scale. Here, our analyses indicate, for instance, that Hungarian students achieved an average of 471 points in RCS-1970 but only 457 points in RLS-1991. Obviously enough, such trends cannot be studied with the *Original_Scores*. In the following sections, we shall evaluate the results from our study more closely. We shall start with a systematic comparison of the original with the new IRT scores. Then we shall take a closer look at long-term trends.

Comparing original and new IRT scores

Earlier in this paper, we compared the assessment material of the different IEA studies considered in the present study. This comparison revealed that they all defined and operationalized reading literacy in a similar way. Therefore, we suggest comparing the rank order of the countries on both scales. Such comparisons, however, are not meaningful across all studies, because the *Original_Scores* are not comparable across assessments. Hence, we examined RCS, RLS, and PIRLS separately.

The PIRLS data revealed basically the same patterns on both reading scales. Hungary, Italy, and the United States achieved roughly the same reading proficiency level in 2001. In comparison to these 3 countries, the Swedish fourth graders achieved a higher and the third graders, a lower proficiency level. In the second and third PIRLS cycles, Hungary, Italy, and Sweden performed quite similarly. The United States showed a somewhat less good result in 2006 and a somewhat better one in 2011. Again, both scales gave the same rank order. Note that both scales also indicated the same trends when we compared the development over the 3 study cycles.

The RLS data are particularly interesting because the *Original_Scores* were based on 3 text types and we disregarded the "documents" in the linking study. Despite this difference, both scales ranked the countries similarly. The only remarkable difference was that the United States performed very well when we looked at the *Original_Scores* from the international reports but somewhat less well on the new *IRT_Scores*. In this context, the international report provides some interesting information on how well the countries performed on the different text types. This information indicates that the United States performed particularly well on the documents in comparison to the other text types (Martin et al., 2003). Insofar it is not surprising that the scores in the international report were favorable for the United States. With regard to the linking study, this finding supports our strategy of excluding the documents because they introduce another source of variance into the achievement tests.

The comparison of the scores in the RCS revealed that the 2 best performing countries in 1970 changed their relative position. The order of the other 2 countries did not change. According to the original international report, Sweden was the best performing country, whereas the new *IRT_Scores* favored Italy. One possible explanation for the observed difference between the scores is that Italian students performed particularly well on the highly discriminating items of the RCS test. Each item contributed equally to the scale when we computed the number of correct items. In IRT, however, the sufficient statistic for trait levels is the sum of raw scores times the item discrimination squared (e.g., Embretson & Reise, 2000; see, for an overview of all item parameters, Appendix C). The differences between the 2 scales are small and we suggest they should not be over interpreted.

Comparing the results from RLS-2001 and PIRLS-2001

The fact that the IEA conducted 2 studies in 2001 provides another opportunity to evaluate the linking study, because we could compare how the countries performed in the 2 studies. In PIRLS-2001, all countries sampled fourth graders and Sweden extended the international study design through an additional third grade sample. In RLS-2001, Italy and the United States sampled Grade 4 while Hungary and Sweden sampled Grade 3. Thus, comparisons were limited to the fourth graders in Italy and the United States and to third graders in Sweden. The z statistic to test whether the means differ can be calculated as follows:

$$z = \frac{\bar{\Theta}_{RLS} - \bar{\Theta}_{PIRLS}}{\sqrt{se^2_{RLS} + se^2_{PIRLS}}} \tag{3}$$

The comparisons revealed basically the same performance level in the Italian RLS and PIRLS samples because the difference was 0 points (rounded) (z = −0.06; p = 0.95). For the United States, the 11-point difference is also not statistically significant (z = −1.41; p = 0.16). A statistically significant difference at the 5% level can be observed for Sweden, where the PIRLS sample outperformed the RLS sample by 8 points (z = 2.33; p = 0.02). However, as the sample sizes for the Swedish samples were particularly large in 2001, we suggest not to overinterpret this finding. The result that different achievement tests point to similar achievement levels provides further empirical support for the assumption that it is possible to link the studies' achievement scales.

International trends over 40 years

Long-term trends in reading literacy at the end of primary school are illustrated in Figure 2. The scores for 2001 are the mean scores from RLS-2001 and PIRLS-2001 weighted by the sample size in the respective studies. The trend lines reveal interesting and diverse patterns. For example, the performance level in the United States has been rather stable over time, but in recent years U.S. students perform somewhat higher. In contrast, a remarkable and continuous decline in performance can be observed in Sweden. Italy outperformed Hungary before the start of the new millennium, but thenceforth students in both countries performed on a same level.

It seems noteworthy that linking adjusts for differences in test difficulty but not for differences in sample composition. Table 3 shows differences in the sample composition in terms of average age and grade. For example, the most recent Swedish samples are composed of somewhat older students from higher grades in comparison to the sample of previous studies. As age and grade are positively associated to student achievement, these differences indicate that the actual decline in performance is even greater as it is depicted in Figure 2. Strietholt et al. (2013) elaborate on the comparability of the samples in international assessments. It is, however, beyond the scope of this study to address this issue in greater detail.

Not-reached items: Incorrect or not-administrated

There are different approaches to calculate test scores if students do not manage to complete a test. In the present study we treat these not-reached items as if they were not administrated because their proportion varied considerably across the different studies (see the section Missing data). An alternative strategy is to treat not-reached items as incorrect responses when student proficiency scores are generated. In this case, the not-reached items are treated as if they were not administrated (= missing) when estimating item parameters but they are considered as incorrect responses when estimating student proficiency.

In order to examine how the different approaches affect our results, we also adopted the second strategy to compute student achievement scores. The correlation between the scores we receive when adopting both approaches is .93 on individual level (student) and .85 on aggregate level (country ×

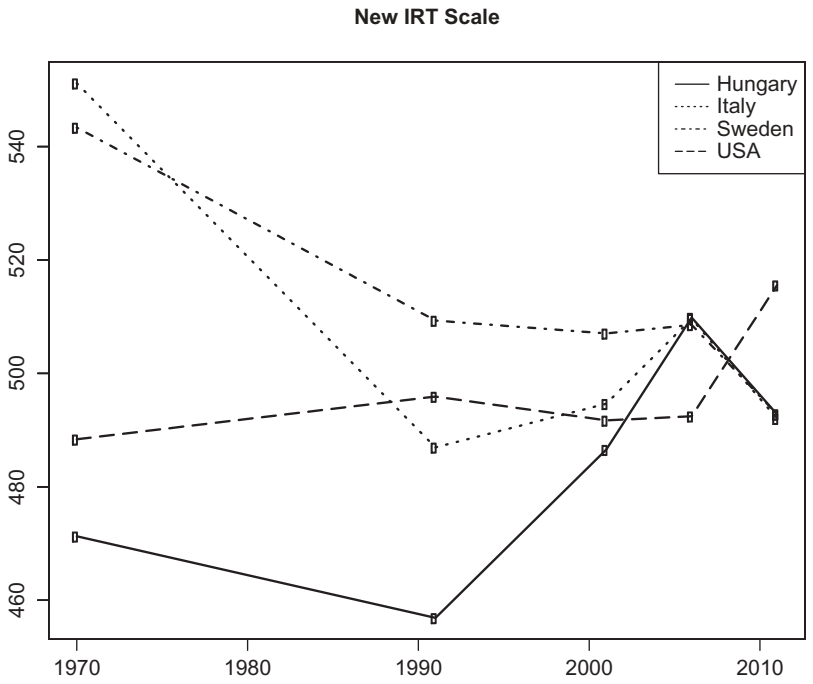


Figure 2. Trends in reading achievement on the common IRT scale.

study). The largest differences between the scores can be observed for samples from RCS-1991/2001. This finding corresponds with the observation that the proportions of not-reached items were particularly large in RCS-1991/2001. As we outlined above, we prefer to treat not-reached items as if they were not administrated because we think that, from a substantive standpoint, this approach produces more comparable scores over time.

Conclusion

Studies such as PIRLS, PISA, and TIMSS receive considerable attention from researchers and policy makers. However, international comparative studies on student achievement are not a recent innovation and have already been conducted for nearly half a century. Combining recent and older data sets opens up the possibility of studying long-term trends in levels of attainment or educational inequalities. Because changes made on, for example, the system level of educational policy typically do not show until the changed structures are fairly well implemented, currently observable outcomes may be due to reforms introduced quite some time in the past. Therefore, long-term-trend studies provide a unique approach to investigate the effects of educational policies on educational outcomes. However, these opportunities have rarely been exploited yet. This may well be because the various studies used different tests and the achievement scores are not comparable over time.

The present study demonstrates how to link the achievement tests from recent and older studies onto the same measurement scale. We employed IRT models and made use of the overlaps in the assessment material in a concurrent calibration of all item parameters onto the same IRT scale. Some of these overlaps are due to national extensions of the international study designs. Furthermore, hence, our study demonstrates how international studies can benefit from national extensions of the study design because the resulting item parameters can be utilized by other countries.

The fact that some overlaps in the assessment material are limited to (Swedish) national extensions also raise questions for future research. First, the current study is based on the pivotal

assumption that the estimation of the item parameters is independent of the sample. Test linking adjusts for differences in test difficulty. The basic idea of a common-item nonequivalent group design is that some students receive items from both Test X and Test Y. The knowledge about how well the same students perform on both tests allows us to link them. However, imagine a situation in which Test X is relatively easy compared to Test Y for Swedish students but both tests are equally difficult in all other countries. In such a situation, the results from the present study would not be trustworthy. This problem is also referred to as differential item functioning (DIF, e.g., Osterlind & Everson, 2009). DIF would be a severe problem for the current study if, for example, the RLS test were easier than the PIRLS test for Swedish students whereas both were equally difficult for students in the other countries. We cannot test for this possibility empirically because the links between some studies are not available in other countries. However, the comparison of the assessment material from the various studies gives no indication of substantial differences. Nonetheless, future research is needed here. For example, different countries could conduct linking studies in which they administer the assessment material from different studies to the same students. Such data could then be used to further explore the robustness of the linking studies. Other methodological issues for future studies are related to the effects of modeling fixed or random item parameters across countries and over time, and examining consequences for (linking) errors (Hsieh, Xu, & von Davier, 2009; Martin, Mullis, Foy, Brossman, & Stanco, 2012; Weeks, von Davier, & Yamamoto, 2013).

Further limitations of the present study are differences in the item formats. For example, the assessment material in the PIRLS studies contained open-structured response items, whereas the older studies used only multiple-choice. We suggest modeling the guessing parameter to deal with this problem. However, one might also argue that open-ended items are problematic because formulating an answer requires specific cognitive capabilities and processes. This then relates to the dimensional structure of a reading test.

Another limitation is the relation between test length and testing time in the respective studies. Pressure of time refers to a speed dimension in achievement tests and its impact may well differ across the studies. Therefore, we suggest treating not-reached items as missing data in order to overcome potential bias in this regard. This strategy, however, punishes students who aim to complete the entire test in comparison to students who spend more time on the items at the beginning. As a result, many studies treat not-reached items as wrong responses. Although we see the advantages and disadvantages of both approaches, we argue that our strategy is more suitable for a trend study using tests that differ in terms of the speed dimension. However, further research is needed to understand the consequences of both approaches from an empirical perspective.

Despite these limitations, our study provides first insights to long-term trends in student reading achievement at the end of primary school. This data enables policy makers and researchers to evaluate long-term developments in educational outcomes. Finally, however, we would like to highlight that the results from the current linking study are merely a starting point for further substantial analyses. To investigate the effects of certain educational policies and policy-related issues, it would be possible to include more countries and future international large-scale studies. For pragmatic reasons, the current study focused on data from countries that participated in all previous IEA studies on reading literacy. Future research may also include countries that did not participate in all studies. Although the issue of missing information may raise further problems here, it seems worth including more countries to increase the sample sizes in more substantive studies. Finally, the present article has not just demonstrated that it is possible to link older and recent large-scale studies. It also indicates that future studies could be linked to this scale as well. The fourth PIRLS cycle will be conducted in 2016. Because the assessment material of this study will be linked to previous PIRLS cycles, it will be possible to add not only more countries but also more time points to the present study.

Acknowledgments

The work is a part of the infrastructure project Center for Comparative Analyses of Educational Achievement (COMPEAT). We wish to thank Laura Zieger for her assistance with the analyses, Thomas Kiefer for the opportunity to use the beta versions of the R package TAM, and Jonathan Harrow for language editing. We would like to thank 2 anonymous referees for helpful comments and discussion.

Competing interests

The authors declare that they have no competing interests.

Funding

This paper was supported by a grant to M. Rosén from the Riksbankens Jubileumsfond (RJ).

Authors' contributions

All authors developed the rationale for the study. The first author analyzed the data and wrote the manuscript. The second author contributed with substantial comments to the manuscript. All authors read and approved the manuscript.

References

Binkley, M. (1994). The IEA reading literacy test. In M. Binkley & K. F. Rust (Eds.), *Reading literacy in the United States: Technical report of the U.S. component of the IEA reading literacy study* (pp. 103–107). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

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

Elley, W. B. (1992). *How in the world do students read? IEA study of reading literacy*. Hamburg, Germany: Grindeldruck.

Elley, W. B. (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems*. Oxford, United Kingdom: Pergamon Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, C. F. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy Press.

Gonzalez, E. J. (2003). Scaling the PIRLS reading assessment data. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2001 technical report* (pp. 151–168). Chestnut Hill, MA: Boston College.

Gonzalez, E. J., Galia, J., & Li, I. (2004). Scaling methods and procedures for the TIMSS 2003 mathematics and science scales. In M. O. Martin, I. V. S. Mullis, & S. J. Chrostowski (Eds.), *TIMSS 2003 technical report: Findings from IEA's trends in international mathematics and science study at the fourth and eighth grades* (pp. 253–273). Boston, MA: TIMSS & PIRLS International Study Center.

Gonzales, E. J., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *3*, 125–156.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal*, *7*(1), 1–17. doi:10.2304/eerj.2008.7.1.1

Gustafsson, J.-E., & Rosén, M. (2006). The dimensional structure of reading assessment tasks in the IEA reading literacy study 1991 and the progress in international reading literacy study 2001. *Educational Research and Evaluation*, *12*(5), 445–468. doi:10.1080/13803610600697179

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Press.

Hanushek, E. A., & Wößmann, L. (2011). The economics of international differences in educational achievement. In E. A. Hanushek, S. Machin, & L. Wößmann (Eds.), *Handbook of the economics of education* (Vol. 3, pp. 89–200). Amsterdam, Netherlands: Elsevier.

Hsieh, C., Xu, X., & von Davier, M. (2009). Variance estimation for NAEP data using a comprehensive resampling-based approach: An application of cognitive diagnostic models. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 161–174.

Joncas, M. (2007). PIRLS 2006 sample design. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2006 technical report* (pp. 35–48). Chestnut Hill, MA: Boston College.

Kiefer T., Robitzsch A., & Wu M. (2015). TAM: Test analysis modules (R package version 1.10–0). Retrieved from http://cran.r-project.org/web/packages/TAM

Kim, S.-H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*(2), 131–143. doi:10.1177/01466216980222003

Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement*, *26*(1), 25–41. doi:10.1177/0146621602026001002

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Lohr, S. L. (2009). *Sampling: Design and analysis*. Boston, MA: Brooke/Cole.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V. S., Foy, P., Brossman, B., & Stanco, G. M. (2012). Estimating linking error in PIRLS. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *5*, 35–47.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., & Kennedy, A. M. (2003). *Trends in children's reading literacy achievement 1991–2001: IEA's repeat in nine countries of the 1991 reading literacy study*. Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (Eds.). (2003). *PIRLS 2001 technical report*. Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 technical report*. Chestnut Hill, MA: Boston College.

Maydeu-Olivares, A. (2013). Goodness-of-Fit Assessment of Item Response Theory Models. *Measurement*, *11*, 71–101.

Mullis, I. V. S., Kennedy, A. M., Martin, M. O., & Sainsbury, M. (Eds.). (2006). *PIRLS 2006. Assessment framework and specifications* (2nd ed.). Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I. V. S., Martin, M. O., Gonzales, E. J., & Foy, P. (2003). *PIRLS 2001 international report: IEA's study of reading literacy achievement in primary schools in 35 countries*. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., & Foy, P. (2007). *PIRLS 2006 international report. IEA's progress in international reading literacy study in primary school in 40 countries*. Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 assessment framework*. Chestnut Hill, MA: Boston College

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*(2), 159–176. doi:10.1177/014662169201600206

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Thousand Oaks, CA: Sage.

Postlethwaite, T. N. (1974). Target populations, sampling, instrument construction and analysis procedures. *Comparative Education Review*, *18*(2), 157–179.

Raîche G., & Magis D. (2015). *Parallel analysis and non graphical solutions to the Cattell Scree Test* (R package version 2.3.3). Retrieved from http://cran.r-project.org/web/packages/nfactors

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non graphical solutions for the Cattell's Scree Test. *Methodology*, *9*(1), 23–29. doi:10.1027/1614-2241/a000051

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. Hoboken, NJ: John Wiley.

Strietholt, R., Rosén, M., & Bos, W. (2013). A correction model for differences in the sample compositions: The degree of comparability as a function of age and schooling. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *6*, pp. 1–20.

Taube, K. (1993). Reading comprehension among Swedish students: A comparative analysis of IEA studies from 1970 and 1991. *Scandinavian Journal of Educational Research*, *37*(1), 89–97. doi:10.1080/0031383930370107

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: An empirical study*. Stockholm, Sweden: Almquist & Wiksell.

Von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9–36.

Wendt, H., Bos, W., & Goy, M. (2011). On applications of Rasch models in international comparative large-scale assessments: A historical review. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *17*(6), 419–446. doi:10.1080/13803611.2011.634582

Weeks, J. P., von Davier, M., & Yamamoto, K. (2013). Design considerations for the program for international student assessment. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 259–275). Boca Raton, FL: Chapman & Hall.

Wolter, K. M. (2007). *Introduction to variance estimation*. New York, NY: Springer.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, *8*, 125–145.

Appendix A. Overview of the Test Passages and Overlaps in the Assessment Material Across Studies.

						RLS		PIRLS		
	Text passage	Text type	Status	No. of items	RCS 1970	1991	2001	2001	2006	2011
1	Poet	LIT	incl.	7	ID					
2	Pole	LIT	incl.	6	ID					
3	Erneke	LIT	incl.	7	ID	SE	SE			
4	The Bird and the Elephant	LIT	incl.	5		ID	ID			
5	Grandpa	LIT	incl.	6		ID	ID	SE		
6	A Shark Makes Friends	LIT	incl.	5		ID	ID	SE		
7	No Dogs Is Not Enough	LIT	incl.	6		ID	ID	SE		
8	The Upside-Down Mice	LIT	incl.	14				ID		
9	Hare Heralds the Earthquake	LIT	excl.[a]	11				ID		
10	The Little Lump of Clay	LIT	incl.	13				ID	ID	
11	Flowers on the Roof	LIT	incl.	13				ID	ID	ID
12	Unbelievable Night	LIT	excl.[a]	12					ID	
13	Fly Eagle	LIT	incl.	12					ID	ID
14	Shiny Straw	LIT	incl.	14					ID	ID
15	The Empty Pot	LIT	incl.	17						ID
16	Enemy Pie	LIT	excl.[a]	15						ID
17	Seal	INF	incl.	5	ID					
18	Ox	INF	incl.	5	ID					
19	Tailor	INF	incl.	5	ID	SE	SE			
20	Plant	INF	incl.	5	ID	SE	SE			
21	Marmot	INF	incl.	4(5)[c]	ID	ID	ID	SE		
22	Postcard	INF	incl.	2		ID	ID	SE		
23	What Is Quicksand?	INF	incl.	3		ID	ID	SE		
24	The Walrus	INF	incl.	6		ID	ID	SE		
25	How to Read the Age of a Tree	INF	incl.	6		ID	ID	SE		
26	River Trail	INF	incl.	11				ID		
27	Night of the Pufflings	INF	excl.[a]	13				ID		
28	Antarctica	INF	incl.	11				ID	ID	
29	Leonardo	INF	incl.	12				ID	ID	ID
30	Searching for Food	INF	excl.[a]	15					ID	

(*Continued*)

Appendix A. (Continued).

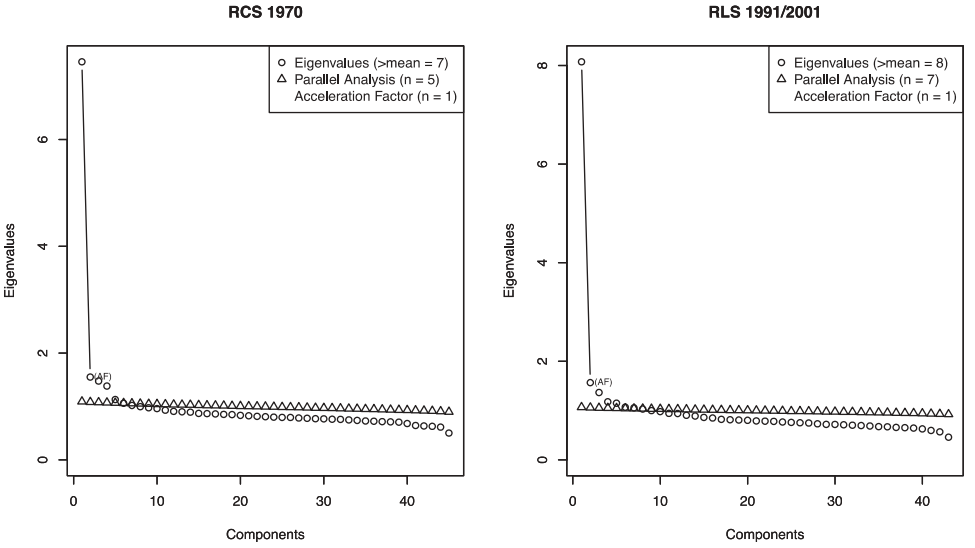
						RLS		PIRLS		
	Text passage	Text type	Status	No. of items	RCS 1970	1991	2001	2001	2006	2011
31	Day Hiking	INF	incl.	12					ID	ID
32	Shark	INF	incl.	12					ID	ID
33	Where's the Honey?	INF	incl.	13						ID
34	The Giant Tooth Mystery	INF	excl.[a]	18						ID
35	Island	DOC	excl.[b]	4		ID	ID			
36	Maria's Timetable	DOC	excl.[b]	3		ID	ID			
37	Buses	DOC	excl.[b]	4		ID	ID	SE		
38	Table of Contents	DOC	excl.[b]	3		ID	ID	SE		
39	Temperature	DOC	excl.[b]	4		ID	ID	SE		
40	Empty Bottles	DOC	excl.[b]	4		ID	ID	SE		

Note. LIT = literacy experience; INF = acquire and use information; DOC = document; ID = link in international design; SE = link in Swedish extension; incl. = included in the present study; excl. = excluded in the present study.

[a]The text passage was excluded because it is part of the PIRLS reader.

[b]The test passage was excluded because it is a document.

[c]The passage Marmot covered five test items in 1970 but only four test items in 1991 and 2006.



Appendix B. Distribution of the eigenvalues for the RCS test and the RLS test.*Note*. The eigenvalues are computed based on data from the international design. We report 3 criteria to determine the number of factors: eigenvalue > 1 (Kaiser-Guttman rule), parallel analysis, and the scree test acceleration factor (AF) (a numerical version of Cattell's elbow criterion; see Raîche, Walls, Magis, Riopel, & Blais, 2006). The *R package nFactors* (Raîche & Magis, 2010) was used for the analyses.

Appendix C. Estimates of the Item Parameters.

		Slope		Location (1)		Location (2)		Location (3)		Guessing	
Text passage	Item	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Tailor	1	1.497	0.027	−2.262	0.027					0.407	0.005
	2	0.738	0.015	−0.536	0.015					0.108	0.011
	3	0.536	0.013	0.172	0.014					0.100	0.009
	4	1.144	0.019	−0.585	0.017					0.040	0.002
	5	0.840	0.015	−0.674	0.015					0.053	0.004
Seal	1	0.966	0.023	−0.747	0.022					0.221	0.008
	2	0.779	0.019	0.068	0.019					0.133	0.011
	3	1.377	0.027	−0.502	0.021					0.152	0.009
	4	0.767	0.018	−1.125	0.020					0.084	0.024
	5	0.536	0.020	0.921	0.023					0.200	0.008
Pole	2	1.628	0.028	−0.945	0.022					0.066	0.005
	3	1.213	0.025	−0.373	0.021					0.138	0.010
	4	0.439	0.018	0.908	0.022					0.175	0.009
	5	0.754	0.018	0.499	0.019					0.066	0.016
	6	0.755	0.018	0.606	0.019					0.062	0.106
	7	0.769	0.023	2.060	0.033					0.177	0.007
Marmot	1	0.932	0.012	−0.634	0.011					0.078	0.003
	2	1.254	0.015	−0.177	0.012					0.104	0.005
	3	0.863	0.012	0.212	0.011					0.111	0.006
	4	0.597	0.017	0.827	0.019					0.072	0.026
	5	0.866	0.012	0.183	0.011					0.147	0.005
Plant	1	1.525	0.023	−1.902	0.022					0.239	0.005
	2	2.151	0.040	−1.457	0.027					0.323	0.005
	3	2.340	0.039	−1.476	0.025					0.295	0.006
	4	1.086	0.019	−0.637	0.017					0.133	0.005
	5	0.734	0.014	0.807	0.015					0.026	0.003
Erneke	1	1.562	0.025	−0.522	0.018					0.121	0.008
	2	1.057	0.017	1.972	0.020					0.093	0.005
	3	1.002	0.017	0.451	0.015					0.071	0.026
	4	0.432	0.013	0.102	0.014					0.068	0.049
	5	0.575	0.021	2.211	0.028					0.080	0.004
	6	0.179	0.014	0.720	0.016					0.162	0.007
	7	0.302	0.012	0.234	0.014					0.082	0.012
Poet	1	1.468	0.029	−0.172	0.023					0.140	0.010
	2	1.531	0.030	−0.506	0.023					0.137	0.010
	3	1.777	0.038	−0.823	0.027					0.138	0.007
	4	1.341	0.027	−0.932	0.023					0.145	0.009
	5	0.728	0.019	0.197	0.020					0.104	0.009
	6	0.422	0.019	1.021	0.023					0.154	0.008
	7	0.877	0.019	0.426	0.019					0.047	0.006

(*Continued*)

Appendix C. (Continued).

		Slope		Location (1)		Location (2)		Location (3)		Guessing	
Text passage	Item	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Ox	1	0.658	0.017	0.402	0.018					0.062	0.148
	2	0.750	0.020	1.283	0.025					0.087	0.007
	3	0.318	0.016	0.502	0.019					0.136	0.010
	4	0.807	0.020	0.353	0.021					0.077	0.006
	5	0.846	0.020	0.173	0.019					0.064	0.021
Postcard	1	0.973	0.016	−2.878	0.023					0.286	0.004
	2	1.017	0.023	−3.688	0.036					0.500	0.004
The Bird and the Elephant	1	1.289	0.018	−0.571	0.015					0.171	0.006
	2	0.814	0.013	−0.908	0.014					0.214	0.006
	3	1.125	0.016	−0.059	0.014					0.118	0.007
	4	1.604	0.020	−2.022	0.020					0.306	0.005
	5	1.476	0.020	−3.131	0.027					0.462	0.004
No Dogs Is Not Enough	1	0.885	0.014	−1.405	0.015					0.134	0.005
	2	1.086	0.016	−0.488	0.014					0.014	0.001
	3	1.267	0.018	−1.705	0.018					0.324	0.006
	4	1.508	0.023	−0.872	0.016					0.203	0.005
	5	1.041	0.017	0.192	0.015					0.130	0.005
	6	1.499	0.022	−1.639	0.019					0.319	0.005
The Walrus	1	0.983	0.019	−2.943	0.027					0.375	0.004
	2	1.217	0.021	−3.309	0.031					0.428	0.004
	3	1.105	0.017	−1.615	0.017					0.178	0.004
	4	0.999	0.016	−1.882	0.018					0.224	0.005
	5	0.752	0.015	−1.556	0.017					0.181	0.006
	6	1.216	0.019	−0.677	0.016					0.148	0.006
What Is Quicksand?	1	0.953	0.015	−0.953	0.014					0.260	0.006
	2	1.473	0.019	−1.667	0.017					0.302	0.004
	3	1.434	0.020	−1.814	0.019					0.317	0.005
A Shark Makes Friends	1	1.673	0.024	−1.947	0.022					0.425	0.004
	2	1.176	0.016	−1.765	0.017					0.127	0.004
	3	0.831	0.014	−0.375	0.013					0.139	0.006
	4	1.338	0.020	−0.830	0.016					0.208	0.005
	5	1.558	0.023	−1.093	0.018					0.375	0.006
Grandpa	1	1.399	0.018	−1.544	0.017					0.163	0.005
	2	2.051	0.025	−1.702	0.020					0.197	0.005
	3	1.199	0.018	−1.077	0.016					0.126	0.005
	4	1.104	0.016	−1.202	0.015					0.098	0.005
	5	2.053	0.030	−1.696	0.023					0.289	0.004
	6	1.197	0.017	−0.941	0.015					0.092	0.010
How to Read the Age of a Tree	1	1.318	0.018	−1.895	0.019					0.276	0.007
	2	0.610	0.013	−0.181	0.014					0.119	0.008

(*Continued*)

Appendix C. (Continued).

		Slope		Location (1)		Location (2)		Location (3)		Guessing	
Text passage	Item	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
	3	1.044	0.019	−0.283	0.016					0.143	0.005
	4	0.783	0.016	0.409	0.015					0.119	0.008
	5	0.850	0.016	−1.534	0.018					0.193	0.008
	6	0.529	0.025	1.910	0.030					0.253	0.006
The Upside-Down	1	1.419	0.053	−2.366	0.057					0.338	0.011
Mice	2	1.605	0.064	−2.924	0.076					0.512	0.010
	3	1.236	0.043	−0.593	0.035					0.097	0.021
	4	0.863	0.031	0.723	0.018						
	5	1.092	0.038	−1.650	0.039					0.135	0.012
	6	1.249	0.025	−2.150	0.010	−0.818	0.013				
	7	1.299	0.037	−1.429	0.011						
	8	0.501	0.027	0.405	0.028					0.000	0.000
	9	1.404	0.050	−1.961	0.049					0.264	0.012
	10	1.115	0.036	−2.699	0.010						
	11	0.831	0.031	−0.245	0.014						
	12	0.781	0.017	−0.982	0.011	0.097	0.015	1.090	0.027		
	13	0.695	0.034	−1.478	0.037					0.061	0.358
	14	0.970	0.034	−1.441	0.011						
The Little Lump	1	1.487	0.031	−1.380	0.009						
of Clay	2	0.863	0.024	−0.278	0.011						
	3	1.323	0.029	−1.940	0.008						
	4	1.006	0.030	−0.795	0.027					0.088	0.022
	5	0.769	0.028	−1.479	0.030					0.061	0.005
	6	1.138	0.028	−1.045	0.009						
	7	1.124	0.036	−1.689	0.035					0.227	0.010
	8	0.832	0.017	−0.552	0.009	0.075	0.013				
	9	1.046	0.031	0.200	0.026					0.095	0.015
	10	0.890	0.015	−1.200	0.008	−0.117	0.010	0.111	0.014		
	11	0.841	0.018	−1.716	0.008	0.718	0.015				
	12	0.617	0.023	−0.851	0.024					0.000	0.004
	13	0.536	0.026	0.351	0.026					0.187	0.012
Flowers on the	1	1.226	0.027	−1.940	0.028					0.185	0.008
Roof	2	0.668	0.021	−1.586	0.024					0.179	0.020
	3	1.037	0.025	−1.771	0.027					0.189	0.008
	4	1.468	0.035	−2.270	0.036					0.362	0.007
	5	1.011	0.021	−1.370	0.022					0.059	0.005
	6	0.876	0.020	−0.842	0.008						
	7	0.662	0.012	0.136	0.008	−0.147	0.011				
	8	1.243	0.023	−1.160	0.007						
	9	1.162	0.015	−2.535	0.006	−1.449	0.007				

(*Continued*)

Appendix C. (Continued).

		Slope		Location (1)		Location (2)		Location (3)		Guessing	
Text passage	Item	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
	10	0.747	0.021	−1.969	0.007						
	11	0.721	0.020	−0.419	0.020					0.160	0.009
	12	0.814	0.013	0.150	0.008	0.205	0.012				
	13	0.947	0.023	−1.213	0.023					0.103	0.009
Antarctica	1	0.999	0.027	−2.196	0.008						
	2	0.891	0.027	−0.645	0.025					0.075	0.033
	3	0.895	0.027	−2.003	0.008						
	4	0.874	0.018	−1.938	0.008	1.044	0.017				
	5	1.311	0.041	−2.767	0.049					0.333	0.009
	6	1.096	0.039	−2.634	0.047					0.317	0.009
	7	1.159	0.016	−2.254	0.008	−1.144	0.009	−0.155	0.012		
	8	0.680	0.017	−2.212	0.008	−0.289	0.011				
	9	0.686	0.017	−1.566	0.008	0.465	0.014				
	10	1.399	0.037	−1.061	0.030					0.077	0.008
	11	0.887	0.025	−0.473	0.010						
Leonardo	1	0.579	0.022	−1.906	0.026					0.120	0.011
	2	0.682	0.021	0.083	0.020					0.132	0.012
	3	0.645	0.018	−0.621	0.008						
	4	1.009	0.012	−1.811	0.007	1.477	0.014	2.029	0.028		
	5	0.952	0.024	0.043	0.021					0.140	0.011
	6	0.700	0.018	−0.112	0.009						
	7	0.652	0.021	0.044	0.020					0.151	0.011
	8	0.917	0.014	−0.762	0.007	1.390	0.016				
	9	1.158	0.025	−1.745	0.026					0.196	0.011
	10	0.903	0.014	−0.004	0.008	0.846	0.014				
	11	0.810	0.021	−1.364	0.022					0.068	0.052
	12	0.722	0.013	−0.631	0.008	1.424	0.018				
River Trail	1	0.713	0.029	−0.963	0.031					0.000	0.004
	2	1.117	0.044	−0.044	0.035					0.106	0.012
	3	1.033	0.040	−1.791	0.043					0.195	0.012
	4	0.945	0.033	−2.159	0.010						
	5	1.154	0.034	−1.903	0.011						
	6	1.024	0.023	−0.931	0.011	−0.785	0.013				
	7	1.171	0.036	−0.319	0.014						
	8	0.983	0.024	−0.946	0.011	0.169	0.017				
	9	0.922	0.023	−0.961	0.011	−0.289	0.015				
	10	0.609	0.017	−1.600	0.010	−0.295	0.014	1.310	0.030		
	11	0.995	0.020	−1.725	0.010	−0.646	0.012	0.418	0.019		
	12	1.012	0.032	−0.565	0.027					0.115	0.015
Shiny Straw	1	1.785	0.054	−1.452	0.038					0.289	0.010

(*Continued*)

Appendix C. (Continued).

		Slope		Location (1)		Location (2)		Location (3)		Guessing	
Text passage	Item	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
	2	0.863	0.025	0.030	0.012						
	3	1.183	0.037	−0.354	0.029					0.171	0.013
	4	1.799	0.083	−0.573	0.052					0.185	0.016
	5	1.693	0.047	−0.833	0.032					0.047	0.003
	6	0.728	0.025	−1.561	0.028					0.046	0.006
	7	1.561	0.043	−1.449	0.035					0.253	0.009
	8	1.162	0.020	−1.847	0.008	−0.723	0.011				
	9	0.890	0.026	0.125	0.013						
	10	1.247	0.035	−1.072	0.029					0.099	0.007
	11	1.055	0.020	−0.453	0.009	−1.063	0.011				
	12	0.978	0.016	−1.239	0.009	0.114	0.012	0.727	0.019		
	13	0.693	0.017	−0.183	0.010	−0.168	0.014				
Fly Eagle	1	1.590	0.048	−2.676	0.052					0.412	0.009
	2	1.242	0.037	−1.453	0.034					0.160	0.010
	3	0.443	0.023	−0.657	0.024					0.004	0.001
	4	1.839	0.057	−2.676	0.058					0.514	0.008
	5	1.089	0.019	−1.676	0.008	−0.891	0.010				
	6	1.602	0.043	−1.336	0.033					0.103	0.006
	7	0.874	0.018	−1.587	0.008	−0.424	0.011				
	8	1.184	0.035	0.366	0.028					0.125	0.013
	9	0.643	0.016	−0.168	0.011	1.024	0.021				
	10	1.317	0.031	−0.730	0.010						
	11	0.825	0.028	−1.003	0.027					0.017	0.003
	12	0.973	0.019	−1.111	0.009	0.231	0.014				
Day Hiking	1	1.053	0.029	−1.591	0.030					0.071	0.006
	2	0.939	0.026	−1.031	0.010						
	3	1.059	0.019	0.687	0.013	1.811	0.027				
	4	1.126	0.033	−0.658	0.028					0.114	0.017
	5	1.741	0.049	−2.689	0.052					0.405	0.009
	6	1.582	0.043	−1.971	0.040					0.272	0.010
	7	1.129	0.032	−1.007	0.028					0.096	0.019
	8	1.012	0.027	−0.767	0.010						
	9	1.057	0.029	−1.526	0.030					0.109	0.017
	10	0.854	0.032	−0.353	0.029					0.149	0.009
	11	0.906	0.018	−0.745	0.009	−0.555	0.012				
	12	0.735	0.017	−0.837	0.009	0.130	0.014				
Shark	1	0.701	0.017	−1.689	0.008	−0.522	0.011				
	2	0.885	0.026	−1.260	0.009						
	3	0.961	0.028	−0.889	0.026					0.013	0.002
	4	0.466	0.027	0.234	0.027					0.091	0.006

(*Continued*)

Appendix C. (Continued).

		Slope		Location (1)		Location (2)		Location (3)		Guessing	
Text passage	Item	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
	5	1.168	0.029	−0.305	0.011						
	6	1.183	0.033	−1.066	0.029					0.130	0.014
	7	0.985	0.019	−0.850	0.009	0.003	0.013				
	8	0.917	0.029	−0.405	0.025					0.062	0.171
	9	1.199	0.034	−1.182	0.030					0.067	0.005
	10	0.968	0.019	0.660	0.012	0.441	0.018				
	11	0.863	0.028	−0.739	0.026					0.072	0.057
	12	0.983	0.016	−1.732	0.008	−0.642	0.010	0.184	0.014		
The Empty Pot	1	1.681	0.061	−2.238	0.058					0.347	0.012
	2	1.159	0.037	−2.263	0.011						
	3	1.441	0.048	−0.971	0.037					0.137	0.019
	4	0.664	0.030	0.315	0.017						
	5	1.555	0.063	−1.355	0.047					0.260	0.013
	6	0.918	0.033	−0.563	0.031					0.062	0.008
	7	1.625	0.059	−1.729	0.052					0.290	0.014
	8	1.729	0.063	−1.629	0.051					0.222	0.011
	9	0.715	0.024	−2.084	0.011	1.167	0.024				
	10	0.600	0.029	0.152	0.016						
	11	1.108	0.045	−1.625	0.046					0.263	0.017
	12	1.075	0.038	−0.747	0.033					0.031	0.006
	13	1.954	0.068	−2.662	0.071					0.450	0.011
	14	2.044	0.070	−1.501	0.050					0.262	0.012
	15	1.099	0.040	−0.947	0.035					0.118	0.020
	16	1.306	0.040	−0.577	0.013						
	17	0.928	0.020	−1.457	0.011	−0.598	0.013	0.277	0.019		
Where's the	1	1.110	0.025	−1.720	0.011	−0.489	0.014				
Honey?	2	1.177	0.027	−0.433	0.012	0.020	0.017				
	3	1.262	0.047	−0.744	0.038					0.137	0.017
	4	0.958	0.034	−1.322	0.012						
	5	0.903	0.037	−0.217	0.032					0.062	0.181
	6	0.731	0.037	−1.666	0.041					0.124	0.016
	7	1.363	0.025	−0.370	0.012	0.276	0.016	0.871	0.022		
	8	1.426	0.051	−1.273	0.042					0.137	0.018
	9	0.924	0.038	0.125	0.033					0.127	0.018
	10	1.282	0.050	−0.097	0.039					0.095	0.014
	11	1.402	0.042	0.491	0.017						
	12	1.194	0.045	0.414	0.037					0.076	0.010
	13	0.916	0.034	0.635	0.019						