The dimensions of reading comprehension in Dutch children: Is differentiation by text and question type necessary?

Muijselaar, M.M.L.; Swart, N.M.; Steenbeek-Planting, E.G.; Droop, M.; Verhoeven, L.; de Jong, P.F.

DOI
10.1037/edu0000120

Publication date
2017

Document Version
Final published version

Published in
Journal of Educational Psychology

Citation for published version (APA):
Muijselaar, M. M. L., Swart, N. M., Steenbeek-Planting, E. G., Droop, M., Verhoeven, L., & de Jong, P. F. (2017). The dimensions of reading comprehension in Dutch children: Is differentiation by text and question type necessary? *Journal of Educational Psychology*, *109*(1), 70-83. https://doi.org/10.1037/edu0000120

Journal of Educational Psychology
2017, Vol. 109, No. 1, 70–83

© 2016 American Psychological Association
0022-0663/17/$12.00 http://dx.doi.org/10.1037/edu0000120

The Dimensions of Reading Comprehension in Dutch Children: Is Differentiation by Text and Question Type Necessary?

Marloes M. L. Muijselaar
University of Amsterdam

Nicole M. Swart, Esther G. Steenbeek-Planting,
Mienke Droop, and Ludo Verhoeven
Radboud University

Peter F. de Jong
University of Amsterdam

Many recent studies have aimed to demonstrate that specific types of reading comprehension depend on different underlying cognitive abilities. In these studies, it is often implicitly assumed that reading comprehension is a multidimensional construct. The general aim of this study was to examine the dimensionality of a large pool of reading comprehension items differing according to text and question type. The items were administered to 996 fourth-grade children. We used multitrait, multimethod modeling to test for the existence of specific text and question types. In addition, the correlations of factor scores, reflecting the different measures of reading comprehension, with word reading speed, vocabulary, and working memory were examined. Confirmatory factor analyses revealed that the specific measures of comprehension, differing according to text and question type, hardly reflected systematic variation, after a general factor of reading comprehension was taken into account. Reading comprehension items thus largely reflect a common factor. Factor scores that were supposed to reflect specific comprehension factors were not reliable and were hardly related to word reading speed, vocabulary, and working memory.

Keywords: reading comprehension, dimensionality, cognitive predictors

A child's level of reading comprehension can be measured with a variety of comprehension tests that differ according to text and question characteristics. A few very early studies suggest that such differences do not represent different aspects of comprehension as the structure of reading comprehension appeared to be merely one-dimensional (Davis, 1944; Spearritt, 1972; Thorndike, 1973). More recently, however, numerous studies have shown that comprehension measures can differ in the comprehension abilities that are assessed and in the cognitive skills that are required (e.g., Andreassen & Bråten, 2010; Bowyer-Crane & Snowling, 2005; Cutting & Scarborough, 2006; Francis et al., 2006; Keenan, Betjemann, & Olson, 2008; Keenan & Meenan, 2014; Kendeou, Papadopoulos, & Spanoudis, 2012; Nation & Snowling, 1997; Spear-Swerling, 2004). In contrast to the early studies, these more recent studies assume that different measures of reading comprehension reflect different subskills. Put differently, they find that the structure of reading comprehension is multidimensional. The main purpose of the current study was to further examine the dimensionality of reading comprehension. The first question was whether it is possible to distinguish specific types of reading comprehension measures in a large item pool. A second question was whether the relations of word reading speed, vocabulary, and working memory with reading comprehension depend on the type of reading comprehension measure. In what follows, first previous research about differences between specific reading comprehension measures is discussed. Then, it is argued why it is necessary to examine the structure of reading comprehension before investigating the relations between cognitive predictors and specific measures of reading comprehension.

Reading Comprehension Measures Differ in Their Cognitive Predictors

Reading comprehension depends on several cognitive processes (Kendeou, van den Broek, Helder, & Karlsson, 2014). According to the Simple View of Reading (Hoover & Gough, 1990), the readers' accuracy in decoding words and linguistic processes, such as vocabulary, are major determinants of reading comprehension (de Jong & van der Leij, 2002; Hoover & Gough, 1990; Tilstra, McMaster, van den Broek, Kendeou, & Rapp, 2009; Verhoeven & van Leeuwe, 2008). In more transparent languages, however, measures of reading accuracy do not discriminate between poor and good readers (Florit & Cain, 2011). In such languages, a speed component is necessary for decoding to be related to reading

This article was published Online First April 28, 2016.

Marloes M. L. Muijselaar, Research Institute of Child Development and Education, University of Amsterdam; Nicole M. Swart, Esther G. Steenbeek-Planting, Mienke Droop, and Ludo Verhoeven, Behavioural Science Institute, Radboud University; Peter F. de Jong, Research Institute of Child Development and Education, University of Amsterdam.

We are grateful to Frans Oort for his statistical advice and to Elise de Bree for proofreading the article. This research was funded by the NWO Programming Council for Educational Research (PROO; 411-10-925).

Correspondence concerning this article should be addressed to Marloes M. L. Muijselaar, Research Institute of Child Development and Education, University of Amsterdam, P. O. Box 15776, NL-1001 NG Amsterdam, The Netherlands. E-mail: m.m.l.muijselaar@uva.nl

comprehension (de Jong & van der Leij, 2002). Beyond decoding and vocabulary, working memory is generally considered an important contributor to reading comprehension as comprehending a text involves the construction of relations between words and sentences and, in the end, the construction of a situation model (e.g., Cain, Oakhill, & Bryant, 2004; Daneman & Merikle, 1996). In the current study, we focused on three cognitive predictors of reading comprehension: word reading speed, receptive vocabulary and working memory.

The involvement of cognitive abilities in reading comprehension tends to vary across reading comprehension tests differing in text and question types. For example, it has been suggested that long texts are more dependent on working memory than short texts (Andreassen & Bråten, 2010). This is line with the construction of a situation model: compared with shorter texts, the comprehension of long texts demands more frequent updates of the situation model and more information has to be stored in memory (e.g., Kintsch, 2012). However, the results on the relations between cognitive processes and particular measures of comprehension tests are equivocal (Andreassen & Bråten, 2010; Basaraba, Yovanoff, Alonzo, & Tindal, 2013; Bowyer-Crane & Snowling, 2005; Cutting & Scarborough, 2006; Eason, Goldberg, Young, Geist, & Cutting, 2012; Francis et al., 2006; Keenan et al., 2008; Keenan & Meenan, 2014; Kendeou et al., 2012; Miller et al., 2014; Nation & Snowling, 1997). For instance, in contrast to the findings by Andreassen and Bråten (2010), that long texts place a larger demand on working memory than short texts, Keenan and Meenan (2014) found the opposite results. Keenan and Meenan take this difference to stem from the format features of the tests; tests with short texts require children to hold more (detailed) information in memory than tests with long texts. Another example of mixed findings on the contribution of cognitive abilities on reading comprehension relates to reading fluency and text genre: García and Cain (2014) argued that reading fluency was more important for narrative texts than for expository texts, whereas the study of Eason et al. (2012) revealed a comparable contribution of fluency to both genres of texts. Although research shows that the contribution of cognitive skills to reading comprehension depends on the reading comprehension test used, it remains unclear which specific text and question types are responsible for the differences in the contributions of the various cognitive skills.

A methodological explanation for the inconsistent findings on the relations between cognitive abilities and specific comprehension tests might be that differences in relations were not always tested for significance (with the exception of Kendeou et al., 2012). For example, regression analyses by Eason et al. (2012) showed that understanding of inferential language had a specific effect on interpretation questions and questions requiring critical analyses and process strategies, but not on initial understanding questions. The differences among the various standardized regression estimates on inferential language were small. For example, the nonsignificant estimate for initial understanding questions was .11, whereas the significant estimates for the other types of comprehension (interpretation and critical analyses/process strategies questions) were .17 and .19, respectively (Table 4 in Eason et al., 2012). These standardized regression estimates might not turn out to be significantly different if tested on significance. More generally, differences in the relationships of cognitive abilities with types of reading comprehension tests might be overestimated if not tested and, in combination with small samples, might prove difficult to replicate.

The inconsistent findings of the relations between cognitive abilities and comprehension measures might also be due to the use of different intact reading comprehension tests (whole tests for reading comprehension with specific text and question characteristics). Such tests could differ in many respects, complicating the interpretation of the findings. For example, Keenan et al. (2008) found differences in the contribution of reading accuracy between comprehension tests with cloze items (gap filling items) and question-and-answer items. However, because these comprehension tests also differed in passage length, an alternative explanation for these differences might be that reading accuracy is more important for short than for long texts. In a few studies, testing of differences among specific comprehension measures was reported. These studies grouped the questions of one reading comprehension test according to text and question characteristics and computed subtest scores (groups of items with a specific text or question characteristic; Basaraba et al., 2013; Eason et al., 2012; Miller et al., 2014). Although the use of subtest scores can be regarded as an improvement over the comparison of intact measures, matching of subtests on (influential) characteristics that are not of interest, might be difficult and only partially successful.

Another problem with the studies that focused on differences between reading comprehension tests is that they implicitly assume that specific reading comprehension measures truly exist. However, these studies did not examine the dimensionality of reading comprehension (Andreassen & Bråten, 2010; Bowyer-Crane & Snowling, 2005; Cutting & Scarborough, 2006; Eason et al., 2012; Francis et al., 2006; Keenan et al., 2008; Keenan & Meenan, 2014; Kendeou et al., 2012; Miller et al., 2014; Nation & Snowling, 1997). Therefore, in the current study the dimensionality of reading comprehension was tested with a set of analyses at the item level. Thereby, an unequal distribution of items over text and question types can be taken into account. Moreover, the analyses provide a more direct test of the existence of specific measures of reading comprehension, as is merely assumed by the a priori formation of subtests.

Examining the Structure of Reading Comprehension

A few early studies that examined the structure of reading comprehension revealed that reading comprehension items mainly reflect a single reading comprehension factor (Davis, 1944; Spearritt, 1972; Thorndike, 1973). Although in these studies exploratory factor analyses of the reading comprehension questions suggested that several factors could be distinguished, reliability analyses revealed that only one of these factors could be considered as reliable. Moreover, the correlations among the different factors were very high. A more recent study tested the existence of literal, inferential, and evaluative factors in a pool of 20 questions originating from one text, while taking into account a general reading comprehension factor (Basaraba et al., 2013). The results of this study showed that, in addition to a general factor, specific reading comprehension factors could also be distinguished thereby providing evidence that reading comprehension is a multidimensional construct. In the current study, more complex structures of items were tested than in those previous studies, because we used 77 items originating from several different texts. When examining the structure of such a large pool of reading comprehension items, the fact that items are nested within several texts should be taken into account. In addition, the possibility that all reading comprehension items are indicators of the same construct should be controlled for with a general reading comprehension factor.

A rigorous method to test the structure of a large pool of reading comprehension items is by using a multitrait, multimethod model (MTMM model; Eid et al., 2008; Maul, 2013). Such a model can be used to separate trait, method, and error components. A trait refers to the construct that is intended to be measured usually by two or more different tests (Little, 2013). The method components represent variance that measures have in common because they entail the same method of measurement. Method variance is often regarded as nuisance variance because it is not of principled interest (Maul, 2013). In the current study, we consider the texts in which the questions are nested as method factors because the measurement of reading comprehension should not depend on the use of particular texts. In contrast, the text and question types are regarded as (specific) trait factors (see Figure 1, Model a, for a simplified illustration of the MTMM model).

The text and question types, or traits, in a MTMM model (see Figure 1, Model a) can be correlated due to relations to a common higher order factor. In a hierarchical factor model, or more specifically, a second-order factor model, the second-order factor represents the relations between the correlated traits (e.g., Gustafsson, 1984, 2002). Such a second-order factor model with separate method factors (see Figure 1, Model b) is more restrictive than a MTMM model and if the model fits, to be preferred over a model with correlated first-order trait factors and method factors only (as in Model 1a; Anthony et al., 2011). A specific form of a second-order factor is the bifactor model, or nested factor model (e.g., Chen, West, & Sousa, 2006; Gustafsson & Åberg-Bengtsson, 2010; Schmid & Leiman, 1957). This model is especially suited to separate general and specific factors, and its interpretation is straightforward. In bifactor models, a general factor represents the variance that all items (or tests) have in common (Chen et al., 2006; Schmid & Leiman, 1957). The uncorrelated specific factors describe the variance that items have in common after the common variance described by the general factor is taken into account. In item bifactor models, each item thus has a loading on the general factor and a second loading on one of the specific factors (Cai, Yang, & Hansen, 2011; Gibbons et al., 2007; Gibbons & Hedeker, 1992; Undheim & Gustafsson, 1987; see Figure 1, Model c). An important advantage of a bifactor model over a second-order factor model is that the variance of a set of items can be decomposed in the variance that is explained by a general reading comprehension factor, and the variance that is explained by the specific factors (Gustafsson & Åberg-Bengtsson, 2010).

Second-order and bifactor models have been regularly used to examine the assumedly hierarchical structure of intelligence (e.g., Carroll, 2003; Gustafsson, 1984, 2002; Undheim & Gustafsson, 1987). More recently, these models have been applied for the description of other cognitive domains, such as phonological awareness, oral language, and literacy (Anthony et al., 2011; Foorman, Koon, Petscher, Mitchell, & Truckenmiller, 2015; Mehta, Foorman, Branum-Martin, & Patrick Taylor, 2005; Papadopoulos, Kendeou, & Spanoudis, 2012). To the best of our knowledge, the study by Basaraba et al. (2013) is the only one that examined the structure of reading comprehension with such complex models. As said, in the study by Basaraba et al. (2013) the structure of a relatively small number of items (20) originating from only one text was examined. A bifactor model was fitted in which each item loaded on a general factor, and on one of the specific factors that represented literal, inferential and evaluative questions. In the current study, items came from different texts and differed with respect to text type and question type. To examine the structure of this pool of items we fitted a model with one general factor (like in a bifactor model) and various specific trait and method factors (like in a MTMM model; see Model c in Figure 1). In this complex MTMM model (see Model c in Figure 1), the indicators are the reading comprehension items, the method factors are represented by the different texts. Specific trait factors are the text and question types, and the general trait is a general reading comprehension factor. As in a bifactor model, all latent factors are specified to be uncorrelated. As a result, each item can be described by its relation with the general factor, with several text and question type factors, and with one of the text factors.

Aims of the Current Study

The general aim of the current study was to examine the dimensionality of reading comprehension. This was tested with several confirmatory factor models (i.e., a one-factor model, a bifactor model, and several MTMM models). In line with the study of Basaraba et al. (2013), we hypothesized that specific text and question dimensions of reading comprehension could be distinguished. After specific dimensions of reading comprehension were determined, we examined the relations among word reading speed, receptive vocabulary, and working memory, and the various specific comprehension measures (i.e., text and question types). Although situation model theory would expect long texts to depend more on working memory than short texts, findings so far have not been consistent and hypotheses of these studies mainly concern differences between reading comprehension tests instead of differences between specific text and question types (Andreassen & Bråten, 2010; Basaraba et al., 2013; Bowyer-Crane & Snowling, 2005; Cutting & Scarborough, 2006; Eason et al., 2012; Francis et al., 2006; Keenan et al., 2008; Keenan & Meenan, 2014; Kendeou et al., 2012; Miller et al., 2014; Nation & Snowling, 1997). Therefore, although we expected that the relations of cognitive abilities with reading comprehension are dependent on the text and question types, previous results give little guidance for specific predictions. The current study might give more information about how text and question types determine the relation of cognitive abilities with reading comprehension.

Method

Participants

Participants were 996 fourth-grade children from 43 Dutch classes of 35 elementary schools. These children participated in a longitudinal intervention study, but for the current study, we only used the pretest data of this study. The sample of schools in this study was heterogeneous with respect to location, percentage of immigrants, and the average level of education of the parents, representing the differences between schools in the Netherlands. The sample consisted of 506 boys and 490 girls with a mean age

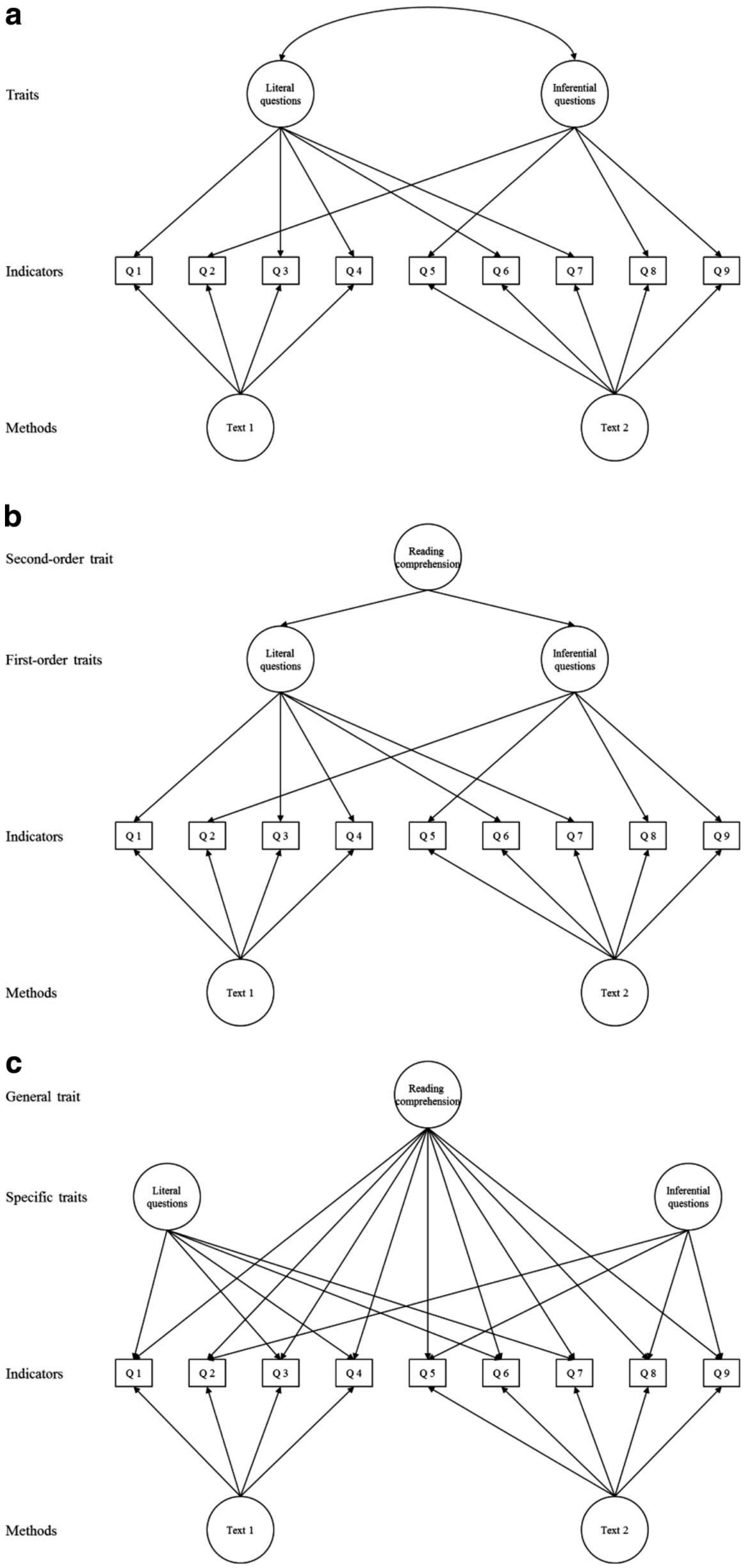

Figure 1. Examples of multitrait, multimethod models for reading comprehension. (a) A model with two correlated trait factors and unrelated method factors. (b) A second-order factor model for the traits and unrelated method factors. (c) A nested factor model with one general trait factor and unrelated method and specific trait factors.

of 9 years and 7 months (*SD* = 5.69 months). Almost 5% of the entire sample was born outside the Netherlands and from 10% of the entire sample, both parents were born outside the Netherlands.

In the Netherlands, children are in elementary school from the age of 4 until the age of 12 years (2 years of kindergarten and Grade 1 through Grade 6). Literacy instruction starts in first grade. Most fourth graders are able to read fluently. Education in reading comprehension usually starts at the end of Grade 2 or at the beginning of Grade 3. Teachers are required to spend 1 to 2 hr per week on teaching reading comprehension. During the comprehension lessons, children are taught to pay attention to important characteristics of a text, such as the title and headings, and also to connectives and linking words. In addition, they learn strategies for how to deal with different texts, such as predicting, questioning, and summarizing.

Design

For the measurement of reading comprehension, a total of 77 reading comprehension questions was used. The questions originated from three different reading comprehension tests (further described in the Instruments section) encompassing nine different texts. The texts differed in text and question types. Text types concerned text genre (narrative or expository) and text length (short or long). Question types were level of comprehension (literal, inferential, and evaluative) and question format (four-option, open-ended, and true-false). Narrative texts had characters and a plot, consist of everyday vocabulary, followed a timeline, were written in past tense, and were often fictional (Best, Floyd, & McNamara, 2008; Eason et al., 2012). Expository texts provided information about a specific topic, included technical vocabulary, and were not structured in a temporal sequence. Literal questions examined children's understanding of information stated explicitly in the text (Basaraba et al., 2013; Eason et al., 2012; Miller et al., 2014). Inferential questions assessed children's ability to make inferences and to draw conclusions about information that was stated in the text. Evaluative questions required an integration between information stated in the text and background knowledge, or required the use of reading strategies; children had to evaluate the information acquired from the text.

The coding of all items with respect to the text and question types was done by the first and second author of this study based on the guidelines provided above. The raters coded the texts and questions independently. There were no differences between the judgments of the text genres. Texts from the Aarnoutse and Kapinga (AK)-Reading Comprehension test with a length of 122 to 288 words were coded as short texts and the texts from the Progress in International Reading Literacy Study (PIRLS) tests with 832 and 920 words were coded as long texts. With respect to level of comprehension (i.e., literal, inferential, and evaluative questions), 73% of the questions were scored similarly. This corresponded with a Cohen's Kappa of .58, which can be classified as a moderate interrater reliability (Viera & Garrett, 2005). The two raters reached consensus about the classification of the other 27% of the questions through discussion. For question format, four-option, true-false questions, and open-ended questions were distinguished. The distribution of the questions over the text and question types is presented in Table 1.

Instruments

Reading comprehension. Reading comprehension was assessed through two different reading comprehension tests. The PIRLS (Mullis, Martin, Gonzalez, & Kennedy, 2003) tests were several reading comprehension tests that contained a narrative or expository text followed by a number of questions. In the current study, one test with a narrative text ("Enemy Pie") and one test with an expository text ("The Mystery of the Giant Tooth") were used. Within each test, four different levels of comprehension were assessed and used to test children's ability to (a) focus on and retrieve explicitly stated information, (b) make straightforward inferences, (c) interpret and integrate ideas and information, and (d) examine and evaluate information in the text. Each test contained two different question formats: multiple-choice and open-ended questions. The multiple-choice questions consisted of four options from which children had to select the correct one. For the open-ended questions, children were asked to write down their answer. Children's answers on the open-ended questions were scored by trained test assistants based on standardized scoring guidelines. Each correct multiple-choice question was awarded 1 point; each (partly) correct open-ended question (1 or) 2 points. The text "Enemy Pie" consisted of 832 words and 16 questions. The text "The Mystery of the Giant Tooth" contained 920 words and 17 questions. Before the start of the test, children were shown examples of how to answer the different question formats. After that, children were asked to read the texts silently and to complete all questions. The texts were available throughout the entire assessment. All children received enough time to finish the test; each text took approximately 40 min to complete. Cronbach's alphas for

Table 1
Distribution of Questions per Text and Question Type

Text/Question type	1a	1b	2a	2b	3a	3b	3c	4a	4b	4c	Total
1a. Narrative texts	—	—									34
1b. Expository texts	—	—									43
2a. Short texts	18	26	—	—							44
2b. Long texts	16	17	—	—							33
3a. Literal questions	8	25	19	14	—	—	—				33
3b. Inferential questions	17	12	14	15	—	—	—				29
3c. Evaluative questions	9	6	11	4	—	—	—				15
4a. Four-option questions	16	21	22	15	13	15	9	—	—	—	37
4b. Open-ended questions	9	9	0	18	7	8	3	—	—	—	18
4c. True-false questions	9	13	22	0	13	6	3	—	—	—	22

the PIRLS test with a narrative and an expository text were .77 and .76, respectively.

The AK-Reading Comprehension test (Aarnoutse & Kapinga, 2006) is part of a standardized battery of tests to measure reading comprehension from Grades 1 through 6. In this study, the test for Grades 4, 5, and 6 was used. The test consisted of a booklet with seven short texts (122 to 288 words) and 44 multiple-choice questions, covering both narrative and expository texts. The multiple-choice questions had either four (A, B, C, D) or two (true-false) options. Each text was followed by six or seven questions: three or four four-option questions and three or four true-false questions. Before the test, one example text was given as a practice trial. Children were required to read all texts silently and complete all questions. All texts were continued to be available during the test. Test administration took approximately 50 min. Cronbach's alpha was .79.

The decoding and comprehension levels of the reading comprehension texts were calculated with the Programma voor berekening Cito LeesIndex voor het Basisonderwijs [Program for calculating the Cito Reading Index for primary education (P-CLIB program; Evers, 2008)]. This program positions levels of text decoding and comprehension to grade levels in which these text levels are generally presented. The decoding level of texts was examined based on average length of the words, and proportion of high-frequent words. The comprehension level was based on the average word length, average sentence length, variation in words, and proportion of high-frequent words. The decoding levels of the nine different texts ranged from halfway Grade 3 to halfway Grade 6. The comprehension levels of the children were between the end of fourth grade and the end of sixth grade.

Word reading speed. For word reading speed we used the *Eén-minuut-test* (Brus & Voeten, 1979). This is a standardized Dutch test often used to measure word reading speed. Children were presented with a list of words of increasing difficulty and asked to accurately read aloud as many words as possible within 1 min. The list consisted of 116 words that increased in length from one to five syllables. The score was the number of words read correctly within 1 min. Reliability scores could not be computed with the data of the present study. The mean parallel-test reliability is .90 (van den Bos, Lutje Spelberg, Scheepstra, & de Vries, 1994).

Receptive vocabulary. An adapted form of the Dutch version (Schlichting, 2005) of the Peabody Picture Vocabulary Test was used to measure receptive vocabulary (Dunn & Dunn, 1997). In the present study, Sets 8 to 13 were used, which consisted of 72 items in total. Each item consisted of four pictures. The test was administered in a classroom setting instead of individually for practical reasons and took approximately 30 min. Children received a booklet with the items and were instructed to underline a picture out of four alternatives that corresponded to the word said by the test assistant. Before the start of the test, two practice items were given. All children finished the entire test. The total score was the number of correct answers. Within our sample, Cronbach's alpha was .69.

Verbal working memory. An experimental listening span test was chosen to measure verbal working memory. For each item, children were required to listen to a series of sentences. The sentences consisted of three to seven words and were presented by a test assistant. After each sentence was presented, the children had to decide whether the sentence was correct and remember the last word of the sentence. The words that had to be remembered were monosyllabic and commonly known by 6-year-old children (Schaerlaekens, Kohnstamm, & Lejaegere, 1999). At the end of each series of sentences, the last words had to be recalled in the same order as the sentences were presented. The test items increased in length from two to five sentences. There were 20 items, four for each number of sentences. The test was administered individually and ended when children failed on all four items of the same number of sentences. Before the start of the test, two example items of respectively one and two sentences were given. The score was the number of items (sentence series) recalled in the correct order. Since this was an experimental test that was stopped when children made too many errors, the reliability could be calculated if the missing items were coded as incorrect. In case of a stopping rule, the difficulty of items is presumed to increase. It can be assumed that these items have been made incorrectly. Based on these assumptions, Cronbach's alpha was .69.

Procedure

The tests were administered in four test sessions. In the first test session, both PIRLS tests were carried out in a classroom setting. On the second day, the AK-Reading Comprehension test and the Peabody Picture Vocabulary test were administered, also in a classroom setting. The listening span and the word reading test were administered individually during a third session.

Analyses

Four questions of the PIRLS tests on which 2 points could be acquired, were recoded to make the scoring of all questions comparable, thus dichotomous. For one easy question (for which 47% of the children had 2 points), both 0 and 1 point were scored as 0, and 2 points was scored as 1 point. For the three more difficult questions (for which 54% to 77% had 0 points), both 1 and 2 points were scored as 1 point.

The structure of the reading comprehension items was examined with several confirmatory factor models. First, a one-factor model was estimated in which all items load on a single reading comprehension factor. Since each item also pertains to one of the nine texts, second, a bifactor model was specified by adding nine text factors (one for the items of each text). Third, different complex MTMM models were estimated. These models are more complex than a standard MTMM model, because all items load on a general trait factor, a method factor and several specific trait factors (see Figure 1, Model c). The factors for text genre, text length, level of comprehension, and question format were added to the model with the general factor and the text factors separately. In the fourth, fifth, and sixth step the text and question type factors were added one by one. In the final model, items loaded on a general reading comprehension factor, one of the nine text factors, and on all text and question type factors (see Figure 2). In this complex MTMM model, the indicators are the reading comprehension items and the methods factors are represented by the different texts. Specific trait factors are the text and question types, and the general trait is a general reading comprehension factor. As in a bifactor model, all latent factors are specified to be uncorrelated. As a result, each item can be described by its relation with the general factor, with several text and question type factors, and with one of the text factors. An alternative model was presented as well, that is, a model without the nine text factors.

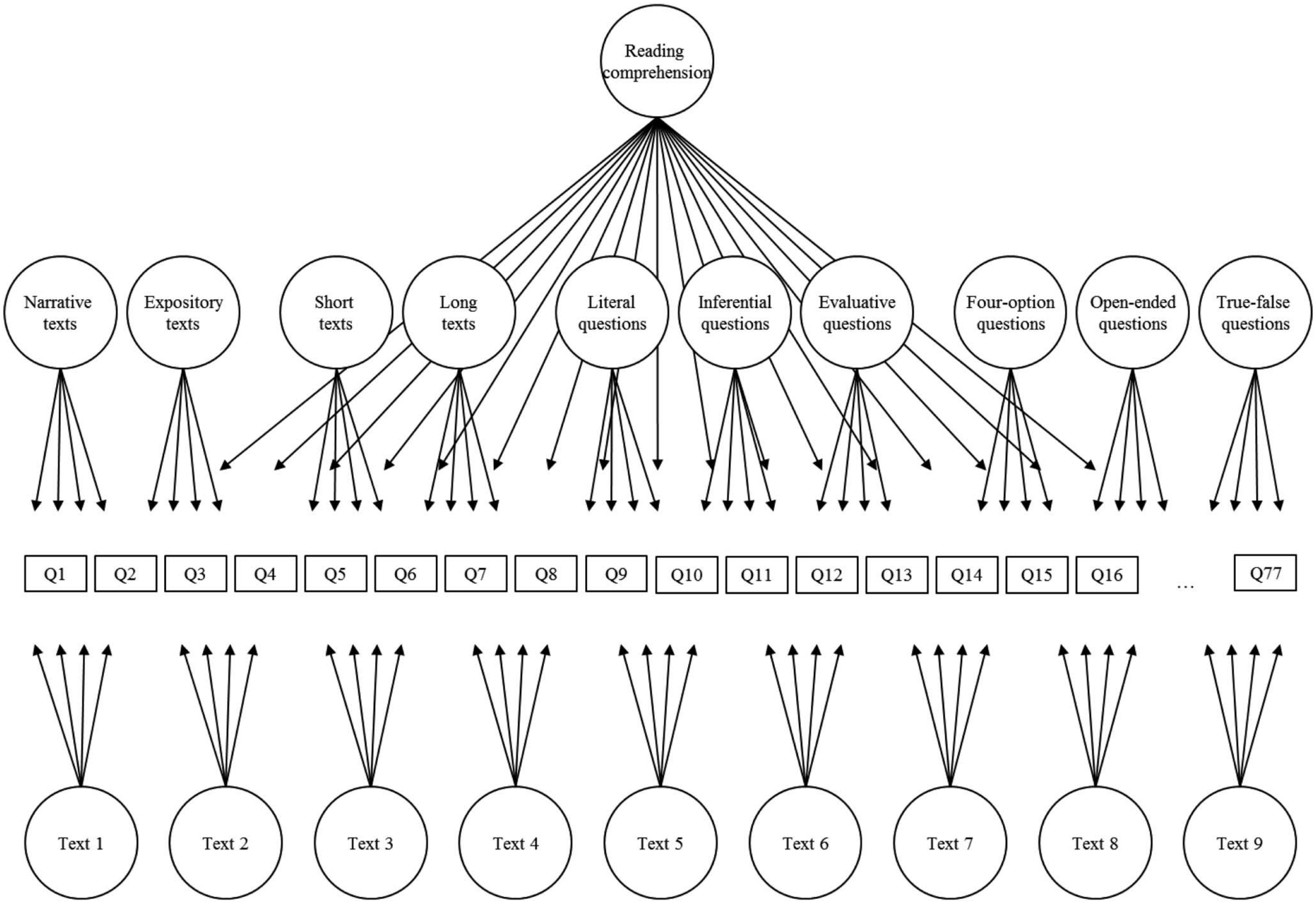

Figure 2. The final complex multitrait, multimethod model. In this model, all items have a loading on the reading comprehension factor (general trait), on four of the text and question type factors (specific traits), and on one of the text factors (methods).

The factor analyses were conducted with *Mplus* Version 7.11 (Muthén & Muthén, 2012). Robust weighted least squares (WLS) estimation was used to obtain parameter estimates. The models only contained dichotomous items and WLS was the estimator; therefore, theta-parameterization was used. Since the children in our sample were nested within classes, there is some dependency in the data. The intraclass correlation coefficients were .08, .12, and .13 for the three different reading comprehension tests. *Mplus* can account for the nested structure of the data and adjust the standard errors accordingly (by using the TYPE = COMPLEX command).

Overall model fit was evaluated with the chi-square goodness-of-fit test-statistic, the root mean square error of approximation (RMSEA), and the comparative fit index (CFI; Kline, 2011). A significant chi-square indicated poor model fit, and a model with a nonsignificant chi-square has good fit to the data. An RMSEA below .05 was taken as good approximate fit, values between .05 and .08 indicated satisfactory approximate fit, and an RMSEA over .10 was considered as poor approximate fit (Browne & Cudeck, 1993). A CFI larger than .95 indicated good incremental model fit, and larger than .90 was considered acceptable (Hu & Bentler, 1999). To test differences in model fit between two nested models, the chi-square difference test was used (Kline, 2011). Because the difference between the chi-square values of two nested models estimated with WLS does not have a chi-square distribution, the regular chi-square difference test is not valid. Therefore, the corrected chi-square difference test (with Satorra-Bentler correction; DIFFTEST option in *Mplus*), which can be calculated with *Mplus*, was used in this study. In addition to the more global model fit, the local fit of the model was investigated by inspecting the factor loadings and calculating reliability scores for the specific factors.

Factor scores for the specific comprehension measures were extracted from the final model to determine whether the relations between reading comprehension and word reading speed, vocabulary, and working memory are dependent on the type of reading comprehension measure. Therefore, the factor scores of the latent factors of the model were added to a dataset with the cognitive predictors. The correlations among those factor scores and word reading speed, vocabulary, and working memory were examined.

Results

Data Screening and Descriptive Statistics

Data were checked for outliers and missing values. Scores that were more than three standard deviations above or below the mean were omitted. In total, less than 1% of the scores were missing. In most cases these were caused by illness of the child.

The maximum, mean, and standard deviation of the measures for reading comprehension, word reading speed, vocabulary, and working memory are displayed in Table 2. All variables were normally distributed with values of skewness ranging from −.56 to .23 and values of kurtosis between −.68 and .62 (Kline, 2011). The correlations of the reading comprehension tests with word reading speed, vocabulary, and working memory were moderate (see Table 3). According to the grade-referenced norms of the AK-Reading Comprehension test and the word reading test, the children in our sample had average levels of word reading and reading comprehension. Therefore, the level of word reading and the comprehension levels of the texts were adequately matched to the ability level of the children. This match is also visible in the fact that floor and ceiling effects were not found on any of the tests.

Testing the Structure of Reading Comprehension

We examined the structure of reading comprehension with a series of confirmatory factor models. As a first step, a model with one general factor was estimated (Model 1 in Table 4). The chi-square value was significant, which indicates poor overall model fit. However, the approximate fit (RMSEA) of this model to the data was good, and the incremental fit (CFI) was acceptable.

Next, models with method factors were specified. In Model 2, a bifactor model was estimated in which a general reading comprehension factor and nine text factors were presumed, one for each text in which the questions were nested. This model could not be estimated. The factor loadings of three items on the corresponding text factor appeared to be extremely high. Fixing these factor loadings to .90 and the residual variances of these items to .19 solved the estimation problems. Both the overall model fit, the approximate fit, and incremental fit of this model were good. In addition, the fit of this bifactor model was significantly better than the fit of the model with a general factor only (chi-square difference test for Model 1 vs. Model 2, see Table 4).

Third, we estimated models in which both method factors and specific reading comprehension factors were included. In these complex MTMM models we specified a general reading comprehension factor, the texts as method factors, and one text or question type as specific trait factors (see Model c in Figure 1, and Models 3a to 3d in Table 4). The overall, approximate, and incremental fit of these models was good, and the fit of all these models was significantly better than the fit of the bifactor model, Model 2 (chi-square difference test for Models 3a–d vs. Model 2, see Table 4).

Table 2
Descriptive Statistics for Reading Comprehension, Word Reading Speed, Vocabulary, and Working Memory Measures

Measure	*N*	Maximum	*M*	*SD*
PIRLS “Enemy Pie”	995	19	10.18	3.35
PIRLS “Mystery of the Giant Tooth”	992	18	8.31	3.52
AK-Reading Comprehension Test	986	44	26.83	6.26
Word reading speed	991	116	61.97	13.38
Peabody Picture Vocabulary Test	988	72	35.15	6.29
Listening span	989	16	5.42	2.36

Note. PIRLS = Progress in International Reading Literacy Study; AK = Aarnoutse and Kapinga.

In Step 3a and the fourth, fifth, and sixth step (Models 3a, 4, 5, and 6 in Table 4), the factors for text genre, text level, level of comprehension, and question format were added step by step. The fit of all models was good and significantly better than the fit of the previous model (chi-square difference test for Model 3a vs. 4, 4 vs. 5, and 5 vs. 6 in Table 4). The fit of the sixth model, including all text and question types, had the best fit. Hence, the MTMM model, with a general reading comprehension factor, nine text factors, four text type factors, and six question type factors, was taken as the final model (see Figure 2 for an illustration of the final model).

Interpretation of the Factor Models of Reading Comprehension

The median, minimum, and maximum factor loadings per specific latent factor are shown in Table 5. The factor loadings of the items on the specific latent factors were often very small or, in some instances, even negative. These relatively low factor loadings show that the items have little in common after controlling for the general reading comprehension factor. The variance explained by the latent factors of the final model was calculated with the following formula: $R^2 = (\Sigma\lambda_i)/77$, where 77 is the total number of items and λ_i is the standardized factor loading of a particular item on a specific latent factor. The general reading comprehension factor explained 18.70% of the variance. The additional variance explained by the text factors, and the text and question type factors, ranged from 0.46% to 2.25%. The low factor loadings of the items and the little additional variance explained by these factors implies that they are hard to interpret.

Additionally, we calculated the reliability of the factor scores that can be derived for each factor from the final model. The reliability of a factor score was calculated with the following formula (Brown, 1989): $\rho_c = (\Sigma\lambda_i)^2/[(\Sigma\lambda_i)^2 + \Sigma\theta_{\varepsilon i}]$. In this formula, ρ_c represents the reliability of the composite or latent factor, λ_i is the standardized factor loading of a particular item on a specific latent factor, and $\theta_{\varepsilon i}$ is the standardized residual variance of an item. Because the residual variances were not provided by *Mplus*, these were calculated with the following formula: $\theta_{\varepsilon i} = 1 - \lambda_i^2$. The general factor score had a reliability of $\rho_c = .94$. The reliabilities of the latent text factor scores ranged from $\rho_c = .00$ to $\rho_c = .48$ (median$_\rho$ = .25; see Table 5). The reliabilities of the latent text and question type factor scores were between $\rho_c = .00$ and $\rho_c = .53$ (median$_\rho$ = .08). In all, these results showed that the reliability of the factor scores derived from the general factor was high, whereas the reliabilities of the factor scores from the specific latent text, text type, and question type factors were low. The reliabilities of the text factor scores were somewhat higher than the reliabilities of the text and question type factor scores.

The low median factor loadings and the subsequent limited additional variance explained by the text and question type factors suggest that the specific trait factors add little to the model when controlling for a general reading comprehension factor and text factors. However, the negative factor loadings might also be explained by overparameterization of the model. To diminish the chance of overfitting and to test whether the specific text and question type factors could explain more variance, we tested an

Table 3
Pearson Correlations Between Reading Comprehension, Word Reading Speed, Vocabulary, and Working Memory Measures

Measure	1	2	3	4	5	6
1. PIRLS "Enemy Pie"	1					
2. PIRLS "Mystery of the Giant Tooth"	.65**	1				
3. AK-Reading Comprehension Test	.62**	.65**	1			
4. Word reading speed	.35**	.40**	.35**	1		
5. Peabody Picture Vocabulary Test	.41**	.45**	.51**	.24**	1	
6. Listening span	.31**	.31**	.33**	.21**	.20**	1

Note. PIRLS = Progress in International Reading Literacy Study; AK = Aarnoutse and Kapinga.
** $p < .01$.

alternative model in which the nine text factors were not included. Thus, this model consisted of a general reading comprehension factor and several specific trait factors (i.e., the specific text and question types). This model had a good fit to the data, $\chi^2(2{,}541) = 2{,}597.92$, $p = .211$, RMSEA = .005, 90% confidence interval (CI) [.000, .008], CFI = .99. Inspection of the model parameters revealed that the median factor loadings of the text and question type factors in this alternative model ranged from −.09 to .21 (median $\lambda = .02$) and the reliabilities of the text and question type factors were between .00 and .63 (median $\rho = .09$). Thus, discarding the text factors and thereby maximizing the variance to be explained by the specific text and question type factors, did not make a difference. Also in this model, factor loadings on the specific factors were low or even negative, making it difficult to denote a particular interpretation to these factors.

Correlations of Cognitive Abilities and Specific Measures of Reading Comprehension

To examine whether the correlations between the cognitive abilities and reading comprehension are dependent on the specific measures of reading comprehension, factor scores were extracted from the final model (Model 6 in Table 4). Obviously, correlations between cognitive abilities and unreliable factor scores are expected to be low. Indeed, as shown in Table 6, correlations of the factor scores derived from the text and question type factors, with word reading speed, vocabulary, and working memory were very low. Put differently, we hardly found any specific relations of the cognitive abilities with the various specific measures of reading comprehension after the general factor was taken into account. In contrast, the correlations of the cognitive abilities with the general reading comprehension factor were substantial. The correlations of word reading speed and working memory with the general reading comprehension factor were moderate, .42 and .36, respectively. The correlation of vocabulary with reading comprehension was high (.54).

Discussion

The general aim of the current study was to examine the dimensionality of reading comprehension. The results of our study revealed that reading comprehension questions could largely be represented by a single reading comprehension factor. Specific factors for text and question types were not reliable and explained very little additional variance. The results also showed that there were hardly any differences in the relations of word reading speed, vocabulary, and working

Table 4
Values of Selected Fit Statistics for the Different Confirmatory Factor Models

Number of the model	Name of the model	χ^2	*df*	RMSEA	90% CI	CFI		$\Delta\chi^{2a}$	Δdf
1	1G	3,219.67**	2,849	.011	[.009, .013]	.92	—	—	—
2	1G9T	2,853.14	2,775	.005	[.000, .009]	.98	1 vs. 2	1,069.21**	74
3a	1G9T2G	2,753.69	2,698	.005	[.000, .008]	.99	2 vs. 3a	149.56**	77
3b	1G9T2L	2,759.88	2,698	.005	[.000, .008]	.99	2 vs. 3b	131.30**	77
3c	1G9T3C	2,761.19	2,698	.005	[.000, .008]	.99	2 vs. 3c	136.04**	77
3d	1G9T3F	2,755.26	2,698	.005	[.000, .008]	.99	2 vs. 3d	155.50**	77
4	1G9T2G2L	2,659.53	2,621	.004	[.000, .008]	.99	3a vs. 4	132.32**	77
5	1G9T2G2L3C	2,568.50	2,544	.003	[.000, .008]	1.00	4 vs. 5	128.63**	77
6	1G9T2G2L3C3F	2,479.75	2,467	.002	[.000, .007]	1.00	5 vs. 6	119.02**	77

Note. RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; 1 = one general reading comprehension factor; 2 = one general factor + nine factors for the different texts; 3a = one general factor + nine text factors + two factors for the different text genres; 3b = one general factor + nine text factors + two factors for the different text lengths; 3c = one general factor + nine text factors + three factors for the different levels of comprehension; 3d = one general factor + nine text factors + three factors for the different question formats; 4 = one general factor + nine text factors + two text genre factors + two text length factors; 5 = one general factor + nine text factors + two text genre factors + two text length factors + three level of comprehension factors; 6 = one general factor + nine text factors + two text genre factors + two text length factors + three level of comprehension factors + three question format factors (final model).
[a] Corrected chi-square difference test (Satorra-Bentler correction).
** $p < .01$.

Table 5
Percent of Variance Explained by Latent Factors and Reliabilities of the Latent Factors in the Final Model

Latent factor	*k*	Median λ	Min λ	Max λ	R^2(%)	ρ	Latent factor	*k*	Median λ	Min λ	Max λ	R^2(%)	ρ
General factor	77	.42	.06	.64	18.70	.94	Narrative texts	34	.04	−.29	.56	1.34	.05
Text 1	16	.08	−.38	.23	.71	.00	Expository texts	43	−.02	−.31	.44	1.56	.00
Text 2	17	.18	−.02	.55	1.54	.45	Short texts	44	−.02	−.65	.37	1.71	.01
Text 3	6	.13	−.11	.87	1.18	.23	Long texts	33	.16	−.08	.49	2.25	.53
Text 4	6	.36	.10	.61	1.28	.48	Literal	33	.10	−.13	.39	1.18	.24
Text 5	7	.14	−.06	.81	1.08	.25	Inferential	29	.04	−.24	.43	1.02	.08
Text 6	6	.26	.01	.82	1.19	.39	Evaluative	15	−.02	−.14	.78	1.43	.13
Text 7	6	−.09	−.61	.10	.59	.14	Four-option	37	.00	−.37	.32	1.04	.01
Text 8	6	.14	−.23	.38	.46	.10	Open-ended	18	.10	−.17	.42	.89	.17
Text 9	7	.22	−.07	.89	1.80	.44	True-false	22	.03	−.15	.82	1.21	.08

Note. k = number of items; λ = factor loading; R^2 = variance explained; ρ = reliability.

memory with these specific factors of reading comprehension when a general reading comprehension factor was taken into account.

The structure of reading comprehension was examined with several confirmatory factor models. These models revealed that all text factors as well as all text and question type factors (text genre, text length, level of comprehension and question format) added significantly to the model fit. The final MTMM model for the 77 comprehension questions consisted of a general reading comprehension factor, nine text factors, and 10 specific factors for the various text and question types. Importantly, however, these specific factors cannot be interpreted as a reflection of specific text and question types, which is in line with the very early studies on the dimensions of reading comprehension (Davis, 1944; Spearritt, 1972; Thorndike, 1973). We observed that only a few items had a substantial loading on each text factor. The loadings of all other items that were expected to be indicative of a factor were generally low. Clearly the few items with a substantial loading on a specific factor have something in common, even after the general reading comprehension factor is controlled. This might be due to the common text, or even passage within the text, to which they belong. Possibly these items within a text are related because they depend on common prior knowledge or are related to the same particular aspect of the situation model of the text.

Table 6
Correlations Between Factor Scores for Specific Reading Comprehension Measures With Word Reading Speed, Vocabulary, and Working Memory

Latent factor	Word reading speed	Vocabulary	Working memory
General factor	.42**	.54**	.36**
Text 1	.01	.03	.05
Text 2	.10**	.06	.05
Text 3	−.06	−.06	−.03
Text 4	.07	.11**	.11**
Text 5	−.04	.08*	.02
Text 6	−.02	.02	.02
Text 7	−.01	.03	−.01
Text 8	.02	.02	.03
Text 9	−.02	.01	.00
Narrative texts	.02	−.04	.01
Expository texts	−.00	−.05	−.01
Short texts	.11**	.03	−.00
Long texts	.07*	.06	.10**
Literal	.08**	.10**	.10**
Inferential	−.02	.08**	.05
Evaluative	.05	.06	.02
Four-option	−.08**	.07*	−.01
Open-ended	.06*	.03	.04
True-false	−.07*	−.08*	−.03

* $p < .05$. ** $p < .01$.

The dimensionality of cognitive constructs has been examined for several decades now (e.g., Anthony et al., 2011; Foorman et al., 2015; Gustafsson, 1984, 2002; Mehta et al., 2005; Papadopoulos et al., 2012). In contrast to the very early studies on the dimensionality of reading comprehension (Davis, 1944; Spearritt, 1972; Thorndike, 1973) in which exploratory factor analyses were mainly used, we used confirmatory factor analyses, in particular, MTMM modeling. This MTMM modeling might be regarded as a specific type of hierarchical modeling which has also been used to examine the structure of intelligence (e.g., Carroll, 2003; Gustafsson, 1984, 2002; Undheim & Gustafsson, 1987). The specific type of hierarchical modeling used in this study, enabled us to decompose the variance explained by the general and specific factors, and also compute the reliability of the factor scores (Gustafsson & Åberg-Bengtsson, 2010).

Our findings indicated that the general reading comprehension factor explained only 18.70% of the variance in the questions, showing that there is a lot of (unexplained) item-specific variance. In a single factor model with all 77 items, the general factor explained 20.37% of the variance. Basaraba et al. (2013) constructed both single factor and bifactor models and found that around 31% of the variance was explained by the general factor. Note that in the current study questions from nine different texts load on the general factor, while in the study of Basaraba et al. (2013), who conducted comparable analyses, all 20 questions originated from one text. In additional analyses, we constructed a factor model with questions originating from a single text and then found comparable percentages of variance explained by the general factor (e.g., 28.71% for the 17 questions from the text "The Mystery of the Giant Tooth" and 34.47% for the 16 questions from the text "Enemy Pie"). Thus, the general reading comprehension factor explains more variance if it is not distinguished from text specific variance. Consequently, as the item pool becomes larger and the number of texts increases, less variance will be explained by the general factor.

The inability to find reliable specific factors of reading comprehension might be caused by the fact that as compared to previous studies, we used a relatively homogeneous set of texts and questions (e.g., Keenan et al., 2008). For example, the Peabody Individual Achievement Test and Woodcock–Johnson Passage Comprehension-3 (e.g., Keenan & Meenan, 2014), that are often used reading comprehension tests in the United States, strongly differ from the Dutch reading comprehension tests included in the present study. The Peabody Individual Achievement Test requires children to read a single sentence and choose the correct picture that best expresses the meaning of the sentence after the sentence is removed. In the Woodcock–Johnson Passage Comprehension-3, children are asked to read passages consisting of one or two sentences and fill in the missing word. In contrast to these reading comprehension tests, standardized Dutch reading comprehension tests always consist of paragraphs with several sentences and never require comprehension of a picture. The Dutch tests consist of texts accompanied by questions with different formats and different levels of comprehension. This study showed that for Dutch comprehension tests, most of the variance of the items is explained by a general reading comprehension factor, and specific item type factors do not explain much additional variance. Possibly, differences in the correlates of reading comprehension measures can only be found when comparing comprehension measures that differ more strongly.

Another explanation for the fact that reading comprehension items are largely represented by a single reading comprehension factor might be that there are many abilities that influence children's comprehension of a text (Shanahan, 2014). It seems unlikely that children would apply a certain subset of those abilities to answer specific reading comprehension questions differing in text or question types. A third explanation might have to do with the way reading comprehension tests are constructed (Shanahan, 2014). To end up with a reliable test, items have to be highly correlated with each other. During the development of a reading comprehension test, items that are not correlated highly with other items will be removed from the test. This reduces the chance that a test measures different subskills of reading comprehension. However, it should be noted that though the construction of each reading comprehension test might led to a one-dimensional test, in the current study the comprehension items came from three different tests. Nevertheless, the items reflected mainly one dimension.

With respect to some text and question types, the finding that specific text and question types could not be distinguished might be considered desirable. For example, questions with different question formats should not require different comprehension processes. However, based on previous studies, we expected to find specific factors for literal and inferential questions. Some previous studies have found that literal questions are easier and require the understanding of information that is literally presented in the text, while inferential questions are more difficult and require inference making (e.g., Basaraba et al., 2013). In the present study however, literal and inferential questions could not be distinguished. Additional analyses of the data in this study showed that children's performance was lower on the inferential questions than on the literal questions. Thus, literal and inferential questions depended on similar comprehension abilities, as revealed by confirmatory factors analyses, but literal questions require a lower level of reading comprehension ability than inferential questions. A reason for the differences between the current study and previous research might be that a substantial number of the children in our study was observed to answer both literal and inferential questions by heart, that is without consultation of the text. These students first read the entire text and then answered all associated questions without looking back to the text. As a result, children used their situation model of the text both for answering literal and inferential questions.

Another finding was that the correlations of word reading speed, vocabulary, and working memory with the specific text and question type factor scores were very low and did not differ substantially. This was to be expected given the unreliability of these specific factor scores. Previous studies did find substantial relations between cognitive abilities and specific measures of reading comprehension (e.g., Eason et al., 2012). The difference in the strength of the relations is probably caused by the fact that these studies did not take into account a general reading comprehension factor. In this study, the correlations of word reading speed, working memory and vocabulary with the highly reliable general reading comprehension factor score were substantial. The correlation of vocabulary with the general reading comprehension factor scores could even be qualified as high. These findings are in line with correlations of these abilities with reading comprehension reported in previous studies (e.g., Oakhill & Cain, 2012).

Limitations

The current study has a number of limitations. The first is the relatively low interrater reliability for the scoring of the reading comprehension items according to level of comprehension. Although a Cohen's kappa of .58 is often qualified as a moderate interrater reliability, and thus is sufficient, it might still be considered undesirably low (McHugh, 2012; Viera & Garrett, 2005). However, Cohen's kappa assumes that raters guess marginal proportions of their ratings (McHugh, 2012). This did not seem to have happened. When raters tried to reach consensus on items on which they had disagreed, it became clear that raters had never guessed outcomes but had always made knowledge-based judgments. In studies were guessing is less likely, the percentage of agreement is a better estimate of the interrater reliability. The percentage of agreement of 73% in the current study can be interpreted as strong agreement (LeBreton & Senter, 2008). Nevertheless, 27% of the items were not scored similarly. To examine whether this has influenced the results, we carried out an additional analysis. In this MTMM model, only the items that were coded similarly were used in a model with a general reading comprehension factor, text factors, and text and question type factors. Also in this additional analysis, the specific text and question type factors turned out to be unreliable. Thus, it seems unlike that the relatively low interrater reliability has affected the results of this study.

A second limitation is the relatively low reliability of the vocabulary test. This low reliability is not in line with studies that used the original form of the Peabody Picture Vocabulary test (Dunn & Dunn, 1997). In the current study, the vocabulary test was administered in class. Therefore, all children were administered the same items and, unlike in the original version, the items were less well adapted to the level of the child. The design of our study probably led to administering too few items, resulting in a decrease of the reliability of the test as compared to the original version. A low reliability of the vocabulary measure might have underestimated its relation with reading comprehension. The reliability of the working memory test was also rather low. Nevertheless, the correlations of these cognitive abilities and the general reading comprehension factor were substantial and generally in line with those found in previous studies (e.g., Oakhill & Cain, 2012). Thus the somewhat lower reliabilities of some of the ability measures did probably have a quite small effect on the results of this study.

Implications and Suggestions for Future Research

Our main conclusion that reading comprehension is a single dimension or skill and does not consist of different subskills has important implications for teachers. However, this does not imply that all tests are equally good as a measure of this single dimension. Some reading comprehension tests will be a better indicator of the general reading comprehension factor than other tests, and, consequently, the scores on these tests will be less affected by construct-irrelevant variance (Gustafsson & Åberg-Bengtsson, 2010).

Another implication is that errors on specific reading comprehension questions are not diagnostic for problems with the acquisition of specific subskills. As a result, it might not be necessary to adapt instruction to different subskills (Shanahan, 2014). Instead, teachers could focus more generally on how students can be instructed to comprehend texts. This implication should, however, be considered with some caution. Future studies should reveal whether the results of this study can be generalized. This study contained a relatively homogeneous set of reading comprehension items. Future studies should focus on reading comprehension tests that differ to a larger extent and at reading comprehension tests in which items are deliberately constructed to measure a specific subskill of reading comprehension. In addition, although at fourth grade the development of reading comprehension seems well underway, it might be that as children grow older further specialization of types of reading comprehension might evolve. If that is the case, reading comprehension might become a multidimensional construct when children grow older.

Conclusions

Examining the dimensionality of a large pool of reading comprehension items in a sample of almost 1,000 fourth graders strongly suggests that reading comprehension is a one-dimensional construct. Specific measures of comprehension varying according to text and question type hardly reflected systematic variation. As a result, the cognitive abilities involved in reading comprehension did not depend on text and question type.

References

Aarnoutse, C., & Kapinga, T. (2006). *Begrijpend lezen* [Reading comprehension]. Ridderkerk, The Netherlands: Onderwijs Advisering.

Andreassen, R., & Bråten, I. (2010). Examining the prediction of reading comprehension on different multiple-choice tests. *Journal of Research in Reading, 33,* 263–283. http://dx.doi.org/10.1111/j.1467-9817.2009.01413.x

Anthony, J. L., Williams, J. M., Durán, L. K., Laing Gillam, S., Liang, L., Aghara, R., . . . Landry, S. H. (2011). Spanish phonological awareness: Dimensionality and sequence of development during the preschool and kindergarten years. *Journal of Educational Psychology, 103,* 857–876.

Basaraba, D., Yovanoff, P., Alonzo, J., & Tindal, G. (2013). Examining the structure of reading comprehension: Do literal, inferential, and evaluative comprehension truly exist? *Reading and Writing, 26,* 349–379. http://dx.doi.org/10.1007/s11145-012-9372-9

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology, 29,* 137–164. http://dx.doi.org/10.1080/02702710801963951

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *The British Journal of Educational Psychology, 75,* 189–201. http://dx.doi.org/10.1348/000709904X22674

Brown, R. L. (1989). Using covariance modeling for estimating reliability on scales with ordered polytomous variables. *Educational and Psychological Measurement, 49,* 385–398. http://dx.doi.org/10.1177/0013164489492011

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Brus, B. Th., & Voeten, M. J. M. (1979). *Een-Minuut-Test, vorm A en B: Verantwoording en handleiding* [One-minute-test, Version A and B: Justification and manual]. Lisse, The Netherlands: Swets and Zeitlinger.

Cai, L., Yang, J. S., & Hansen, M. (2011). Generalized full-information item bifactor analysis. *Psychological Methods, 16,* 221–248. http://dx.doi.org/10.1037/a0023350

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology, 96,* 31–42.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports *g* and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence* (pp. 5–22). Oxford, United Kingdom: Pergamon Press. http://dx.doi.org/10.1016/B978-008043793-4/50036-2

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41,* 189–225. http://dx.doi.org/10.1207/s15327906mbr4102_5

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading, 10,* 277–299. http://dx.doi.org/10.1207/s1532799xssr1003_5

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review, 3,* 422–433. http://dx.doi.org/10.3758/BF03214546

Davis, F. B. (1944). Fundamental factors of comprehension in reading. *Psychometrika, 9,* 185–197. http://dx.doi.org/10.1007/BF02288722

de Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. *Scientific Studies of Reading, 6,* 51–77. http://dx.doi.org/10.1207/S1532799XSSR0601_03

Dunn, L. M., & Dunn, L. M. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology, 104,* 515–528. http://dx.doi.org/10.1037/a0027182

Eid, M., Nussbeck, F. W., Geiser, C., Cole, D. A., Gollwitzer, M., & Lischetzke, T. (2008). Structural equation modeling of multitrait-multimethod data: Different models for different types of methods.

Psychological Methods, 13, 230–253. http://dx.doi.org/10.1037/a0013219

Evers, G. (2008). *Programma voor berekening Cito LeesIndex voor het Basisonderwijs. P-CLIB versie 3.0* [Program for calculating the Cito Reading Index in Primary Education. P-CLIB version 3.0]. Arnhem, The Netherlands: Cito.

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies? *Educational Psychology Review, 23,* 553–576. http://dx.doi.org/10.1007/s10648-011-9175-6

Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology, 107,* 884–899. http://dx.doi.org/10.1037/edu0000026

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Auglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10,* 301–322. http://dx.doi.org/10.1207/s1532799xssr1003_6

García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research, 84,* 74–111. http://dx.doi.org/10.3102/0034654313499616

Gibbons, R. D., Darrell Bock, R., Hedeker, D. R., Weiss, D. J., Segawa, E., Bhaumik, D. K., . . . Stover, A. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement, 31,* 4–19. http://dx.doi.org/10.1177/0146621606289485

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika, 57,* 423–436. http://dx.doi.org/10.1007/BF02295430

Gustafsson, J. E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8,* 179–203. http://dx.doi.org/10.1016/0160-2896(84)90008-4

Gustafsson, J. E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73–95). Mahwah, NJ: Erlbaum.

Gustafsson, J. E., & Åberg-Bengtsson, L. (2010). Unidimensionality and interpretability of psychological instruments. In S. E. Embretson (Ed.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 97–121). Washington, DC: American Psychological Association. http://dx.doi.org/10.1037/12074-005

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2,* 127–160. http://dx.doi.org/10.1007/BF00401799

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1–55. http://dx.doi.org/10.1080/10705519909540118

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading, 12,* 281–300. http://dx.doi.org/10.1080/10888430802132279

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of Learning Disabilities, 47,* 125–135. http://dx.doi.org/10.1177/0022219412439326

Kendeou, P., Papadopoulos, T. C., & Spanoudis, G. (2012). Processing demands of reading comprehension tests in young readers. *Learning and Instruction, 22,* 354–367. http://dx.doi.org/10.1016/j.learninstruc.2012.02.001

Kendeou, P., van den Broek, P., Helder, A., & Karlsson, J. (2014). A cognitive view of reading comprehension: Implications for reading difficulties. *Learning Disabilities Research & Practice, 29,* 10–16. http://dx.doi.org/10.1111/ldrp.12025

Kintsch, W. (2012). Psychological models of reading comprehension and their implications for assessment. In J. P. Sabatini, E. R. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 21–38). Plymouth, NH: Rowman and Littlefield Education.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. http://dx.doi.org/10.1177/1094428106296642

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

Maul, A. (2013). Method effects and the meaning of measurement. *Frontiers in Psychology, 4,* 169. http://dx.doi.org/10.3389/fpsyg.2013.00169

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22,* 276–282. http://dx.doi.org/10.11613/BM.2012.031

Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Patrick Taylor, W. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1–4. *Scientific Studies of Reading, 9,* 85–116. http://dx.doi.org/10.1207/s1532799xssr0902_1

Miller, A. C., Davis, N., Gilbert, J. K., Cho, S. J., Toste, J. R., Street, J., & Cutting, L. E. (2014). Novel approaches to examine passage, student, and question effects on reading comprehension. *Learning Disabilities Research & Practice, 29,* 25–35. http://dx.doi.org/10.1111/ldrp.12027

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

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

Nation, K., & Snowling, M. (1997). Assessing reading difficulties: The validity and utility of current measures of reading skill. *The British Journal of Educational Psychology, 67,* 359–370. http://dx.doi.org/10.1111/j.2044-8279.1997.tb01250.x

Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16,* 91–121. http://dx.doi.org/10.1080/10888438.2010.529219

Papadopoulos, T. C., Kendeou, P., & Spanoudis, G. (2012). Investigating the factor structure and measurement invariance of phonological abilities in a sufficiently transparent language. *Journal of Educational Psychology, 104,* 321–336. http://dx.doi.org/10.1037/a0026446

Schaerlaekens, A., Kohnstamm, D., & Lejaegere, M. (1999). *Streeflijst woordenschat voor zesjarigen* [Vocabulary target list for six-year-olds]. Lisse, The Netherlands: Swets and Zeitlinger.

Schlichting, L. (2005). *Peabody picture vocabulary test-III-NL*. Amsterdam, The Netherlands: Harcourt Test Publisher.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22,* 53–61. http://dx.doi.org/10.1007/BF02289209

Shanahan, T. (2014). How and how not to prepare students for the new tests. *The Reading Teacher, 68,* 184–188. http://dx.doi.org/10.1002/trtr.1315

Spearritt, D. (1972). Identification of sub-skills of reading comprehension by maximum likelihood factor analysis. *ETS Research Bulletin Series, 1972,* i-24. http://dx.doi.org/10.1002/j.2333-8504.1972.tb00192.x

Spear-Swerling, L. (2004). Fourth graders' performance on a state-mandated assessment involving two different measures of reading comprehension. *Reading Psychology, 25,* 121–148. http://dx.doi.org/10.1080/02702710490435727

Thorndike, R. L. (1973). *Reading comprehension education in fifteen countries: An empirical study. International Studies in Evaluation III*. Uppsala, Sweden: Almqvist and Wiksell.

Tilstra, J., McMaster, K., van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading, 32,* 383–401. http://dx.doi.org/10.1111/j.1467-9817.2009.01401.x

Undheim, J. O., & Gustafsson, J. E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research, 22,* 149–171. http://dx.doi.org/10.1207/s15327906mbr2202_2

van den Bos, K. P., Lutje Spelberg, H. C., Scheepstra, A. J. M., & de Vries, J. R. (1994). *De Klepel: Een test voor de leesvaardigheid van pseudowoorden* [The Klepel: A test for the reading skills of pseudowords]. Lisse, The Netherlands: Swets and Zeitlinger.

Verhoeven, L., & van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology, 22,* 407–423. http://dx.doi.org/10.1002/acp.1414

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37,* 360–363.

Received February 24, 2015
Revision received February 11, 2016
Accepted February 12, 2016 ■