Proﬁling children’s reading comprehension: A dynamic approach

Sophie Gruhn, Eliane Segers, Jos Keuning, Ludo Verhoeven

Radboud University, Behavioural Science Institute, Postbus 9104, 6500 HE Nijmegen, The Netherlands
University of Twente, Faculty of BMS, Department IST, PO Box 217, 7500 AE Enschede, The Netherlands
Cito, Institute for Educational Measurement, Arnhem, The Netherlands
University of Curacao, Faculty of Arts, Jan Noorduynweg 111, Curacao.

ARTICLE INFO
Keywords:
Reading comprehension
Primary school
Individual differences

ABSTRACT
To proﬁle children’s reading comprehension, we developed a dynamic approach with componential abilities (orthographic knowledge, vocabulary, sentence-integration) being assessed within the same texts and provided with feedback in addition to the global comprehension of these texts. In 275 Dutch third to ﬁfth graders, we investigated to what extent the response accuracy for questions on componential abilities on ﬁrst attempts and after feedback predicted global text comprehension within the same texts as well as the prospective development in a standardized reading comprehension test. We found that global text comprehension was increased by each correctly answered question on a componential ability on ﬁrst attempts and by each correctly answered sentence-integration question after feedback. The accuracy on ﬁrst attempts also explained unique variance of the growth in the standardized reading comprehension test. A dynamic approach may thus help to arrive at a better understanding of the proﬁles of children’s reading comprehension.

1. Proﬁling children’s reading comprehension: A dynamic approach

Reading comprehension is important for educational success (Hakkarainen et al., 2013). Efﬁcient instruction is, thus, highly relevant. Reading comprehension is a complex interaction of word-, sentence-, and text-level processes. This is why comprehension problems have various origins (Perfetti & Stafura, 2014). Individual instructional needs can be derived from a proﬁle on the strengths and weaknesses in the underlying componential abilities in reading comprehension (Cain & Oakhill, 2006). Moreover, examining which children are at risk for a low responsiveness to instruction takes a more preventative approach (Compton et al., 2012; Vaughn & Fuchs, 2003). Standardized reading comprehension tests do not identify the underlying componential abilities in reading comprehension and children’s ability to learn (Cain & Oakhill, 2012; Compton et al., 2012). The alternative of assessing each componential ability with isolated tests does not consider how the components interact and depend on each other (Perfetti & Adlof, 2012; Sabatini et al., 2016). Thus, isolated measures may only be seen as a proxy of interactive reading comprehension processes. A dynamic approach in which the componential abilities (orthographic knowledge, vocabulary, sentence-integration) are assessed within the same texts and the responsiveness to feedback after mistakes is measured may provide a better insight into the required focus and intensity of instruction (Den Ouden et al., 2019). It is unclear yet to what extent this may indeed yield a better understanding of the variation in reading comprehension in perspective of its growth than traditional reading comprehension tests. Therefore, we examined in the present study how the componential abilities of reading comprehension on the ﬁrst attempt and after feedback predict global text comprehension within the same texts and growth in a standardized reading comprehension test.

1.1. Interaction of componential abilities in reading comprehension

Reading comprehension is a complex interaction of several components, which can broadly be separated into lower-order and higher-order abilities (Perfetti, 1999; Perfetti & Stafura, 2014). Lower-order
componential abilities refer to knowledge at the lexical level, and higher-order componential abilities entail competences at sentence and text level. The specificity with which the phonological, orthographic, and semantic information of a word is stored in the mental lexicon and the interconnection strength between each informational node determine the ease of lexical retrieval during reading and, consequently, the availability of cognitive capacities for higher-order processes (Perfetti & Hart, 2002). By this, the quality of lexical representations directly impacts reading comprehension (Richter et al., 2013; Verhoeven & Van Leeuwe, 2008).

Higher-order processes are based on the input from lower-order componential abilities and aim at the construction of meaning beyond single words (Perfetti, 1999; Perfetti & Stafura, 2014). In order to understand the text, the reader has to connect all identified words and phrases to represent the literal text meaning, which is referred to as textbase (Van Dijk & Kintsch, 1983). Since not all information is directly expressed in the text, the reader has to interpret the link between adjacent phrases and sentences (e.g., cohesive ties, semantic mapping) and has to infer implicitly provided information across the whole text by integrating background knowledge (Cain & Oakhill, 2014). Via these so-called sentence-integration abilities, the reader constructs a more abstract and elaborate mental representation of the situation described in the text (situation model), which goes beyond the literal meaning (Van Dijk & Kintsch, 1983). Lower-order and higher-order componential abilities can be seen as interdependent (Cain et al., 2003; Cain & Oakhill, 2014; Daugaard et al., 2017). Several studies found that both uniquely predict reading comprehension abilities (Oakhill & Cain, 2012; Silva & Cain, 2015).

### 1.2. Profiling of individual differences in reading comprehension

The dissociation between decoding and language comprehension has often been used to profile children with reading comprehension difficulties (Aaron, 1991; Kleinz et al., 2017). Children can have a relatively specific problem with either decoding or language comprehension only, or they may perform low on both (Bishop & Snowling, 2004). This is aligned with the simple view of reading that considers reading comprehension as the product of decoding and language comprehension (Gough & Tunmer, 1986). Language comprehension problems may be due to a variety of causes at word, sentence, or text level (problems in vocabulary, (morpho-)syntax, sentence-integration, comprehension monitoring, or story structure knowledge) and, thus, ask for more fine-grained profiling (Clarke et al., 2010; Landi & Ryherd, 2017). Children with comprehension problems might show weaknesses at only one or two levels (Cain & Oakhill, 2006; Colenbrander et al., 2016; Nation et al., 2004). This suggests the existence of various profiles.

Profiles (i.e., performance pattern across word, sentence, and text level) can describe the instructional needs of a child. Therefore, it is important to assess componential abilities in reading comprehension throughout primary school (Sabatini et al., 2014). This may focus on lexical quality (i.e., orthographic knowledge, vocabulary) and sentence-integration because they are key components of reading comprehension, vary among individuals with different reading comprehension abilities, and are responsive to instruction (Ellerman, 2017; Ellemans et al., 2009; Oakhill & Cain, 2012; Perfetti & Stafura, 2014; Therrien, 2004). Such selection criteria of components for assessments were suggested by Perfetti and Adlof (2012).

### 1.3. Problems in current assessment practice

Standardized reading comprehension tests usually measure the final product of comprehension (Kintsch, 2012; Van den Broek, 2012). Such assessments are useful to identify struggling children (Van den Broek, 2012). However, they do not provide information on what leads to weak performances (Cain & Oakhill, 2012; Mislevy & Sabatini, 2012). Furthermore, the results may depend on the choice of tests, as tests vary in their reliance on componential abilities (Colenbrander et al., 2017; Keenan et al., 2008). Therefore, standardized reading comprehension tests should be used with caution when it comes to determining instructional needs. Another limitation in the current practice is that developmental delays are often only identified over time when children’s progress-rate is too flat or stagnates (Compton et al., 2012; Vaughn & Fuchs, 2003). Children first need to fall behind before instructions are intensified, which has been criticized as a wait-to-fail approach. Identifying these children earlier as well as establishing their required intensity of instruction is desirable (Fuchs et al., 2012). To find out what children need to improve on reading comprehension, isolated tests on componential abilities of reading comprehension (e.g., reading proficiency and vocabulary tests) are often administered. However, reading comprehension cannot be considered as just the sum of its parts (Perfetti & Adlof, 2012; Sabatini et al., 2014). Conclusions about higher-order componential abilities are constrained by their dependence on lower-order componential abilities. Vice versa, higher-order componential abilities may be used to compensate difficulties in the fundamental skills (Cain et al., 2003; Nation & Snowling, 1998). In this respect, we should consider reading comprehension as a complex construct with its processes being dependent on the interrelatedness of underlying components. This interrelatedness does not take place in a context-free interplay. The processing behavior is influenced by task demands and text features in interaction with individual differences (Eason et al., 2012; Francis et al., 2018; Wang et al., 2017). Assessing the componential abilities in isolation limits this complexity (Sabatini et al., 2014; Sabatini et al., 2016). Therefore, isolated measures may not be comparable with the actual online processes.

### 1.4. Dynamic approach to reading comprehension assessment

The above articulated problems indicate a need for assessments which provide more information for instruction and consider reading comprehension as an interactive process. The changing view of the complexity of reading comprehension and the need to individually adapt instructions led to more dynamic approaches to assessments (Sabatini et al., 2016; Sabatini et al., 2020).

The first aspect of a dynamic approach is to assess the componential abilities in reading comprehension in a more interactive way. This has been done by measuring the components within the same text (Den Ouden et al., 2019; Sabatini et al., 2015; Sabatini et al., 2016). Sabatini et al. (2015) proposed a reading component battery where two text-level tasks were assessed within the same paragraphs. The word- and sentence-level measures were not integrated to these paragraphs. The performance of sixth graders on each subtask uniquely predicted their score on a reading comprehension test (see Sabatini et al., 2014). This test was based, however, on different texts than the component battery. In contrast, word-, sentence-, and text-level skills were assessed within the same text in another assessment of Sabatini et al. (2016). They showed that the overall performance of kindergarteners to third graders on the assessment was predicted by their prior background knowledge and the knowledge acquired during reading or listening, even after controlling for grade. The assessment was limited, though, to only one text and decoding was not measured despite its relevance for profiling of readers (Bishop & Snowling, 2004; Perfetti, 1999). Across a larger number of texts, Den Ouden et al. (2019) assessed word-, sentence-, and text-level components within each text. They found that the word-level abilities of fourth graders summarized across texts predicted the global comprehension of such texts. It was not examined how the performance at sentence level explained global text comprehension due to a limited reliability of the task.

The second aspect of a dynamic approach is to provide feedback after mistakes during the assessment. Assessments which include forms of instruction are considered dynamic (Sternberg & Grigorenko, 2002). The responsiveness to instruction (e.g., the ability to answer correctly
after feedback) indicates the learning potential. This can estimate the required intensity of instruction during classroom activities and can help identifying children with a low responsiveness to instruction earlier (Fuchs et al., 2012; Gustafson et al., 2014). To date, there are only a few attempts to measure children’s learning potential in reading comprehension by providing instruction during assessment (e.g., Dörfler et al., 2017; Elleman et al., 2011; Navarro & Mora, 2011; Sabatini et al., 2020). To evaluate the benefit of such assessments, it should be considered how much unique variance in educational achievements they explain on top of traditional tests (Caffrey et al., 2008).

The computerized Global-Integrated Scenario-Based Assessment (GISA) examined how K-12 students solve a complex reading task via several underlying subtasks and make use of strategic hints (Sabatini et al., 2020). The total performance was strongly correlated with results on other reading comprehension tests. It has not been examined how the responsiveness to hints was related to achievements at GISA or other test results, and the subtasks rather focused on higher-order componential abilities. Also, Dörfler et al. (2017) and Elleman et al. (2011) focused on only one level of componential abilities. In both studies, inferential hints were provided after incorrectly answered sentence-integration questions. Sixth graders who received person-mediated inferential hints during practice answered similar questions on new texts better than their peers who did not receive hints (Dörfler et al., 2017). Moreover, Elleman et al. (2011) found that the amount of person-mediated feedback required by second graders together with the transfer to new questions explained unique variance in standardized reading comprehension test scores on top of standardized decoding and vocabulary measures. The amount of unique explained variance was, however, not allocated to the accuracy on the first attempt or after feedback.

Others assessed a broader range of componential abilities and provided feedback (Den Ouden et al., 2019; Navarro & Mora, 2011). In Navarro and Mora’s (2011) assessment, pre-determined one-on-one interactions took place between primary- or secondary-education students and an instructor. The latter gave feedback following guidelines and evaluated the students on a checklist. The total assessment score uniquely predicted on top of a standardized reading comprehension test how teachers evaluated the student’s achievements. The unique effect of the responsiveness to feedback was not examined. Besides the assessment being rather time-consuming, the abilities at word and sentence level were not integrated to texts. In contrast, Den Ouden et al. (2019) provided computerized feedback on word-level components within the same texts and related the accuracy on first attempts and after feedback to the global text comprehension of the same texts. They did not find, however, an effect of the accuracy after feedback on global text comprehension over and above the accuracy on first attempts. As summarized scores across texts were used for the analysis, considering the relationships at text level may yield different results.

1.5. The present study

We have articulated that a dynamic approach to reading comprehension assessment at word, sentence, and text level could lead to optimal profiling of children’s instructional needs. In this approach, the componential abilities are assessed within the same text on the first attempt and after feedback has been provided. Existing assessments followed this approach only to a limited extent. They assessed componential abilities only at one or two levels within the same text (Den Ouden et al., 2019; Sabatini et al., 2015) or not over a larger number of texts and without feedback (Sabatini et al., 2016). Others who provided feedback did not do so on componential abilities at several levels (Den Ouden et al., 2019; Dörfler et al., 2017; Elleman et al., 2011; Sabatini et al., 2020) or not on components within the same texts (Navarro & Mora, 2011). It is not yet clear how componential abilities within the same text without and with feedback may be used for profiling of children’s reading comprehension and if such a dynamic approach provides a better understanding of the variation in children’s reading comprehension than traditional reading comprehension tests.

In the present study, Dutch third to fifth graders’ orthographic knowledge, vocabulary knowledge, and sentence-integration abilities were assessed within the same texts on the first attempt and second attempt after feedback has been provided when a mistake was made. Additionally, the global text comprehension of the same text was evaluated. Prior and after this assessment, the standardized reading comprehension test from the Dutch monitoring system has been conducted at schools. The following questions were addressed:

1) To what extent do the responses to the orthographic knowledge, vocabulary, and sentence-integration questions on first attempts and after feedback uniquely predict global text comprehension within the same text?
2) To what extent do the responses to the orthographic knowledge, vocabulary, and sentence-integration questions on first attempts and after feedback uniquely predict the growth in a standardized reading comprehension test?

With respect to the first question, we expected that the probability to answer the global text comprehension questions correctly is increased by each correctly answered componential ability question on the first or second attempt in contrast to an incorrect response. We assumed a larger increase for correct first than second attempts. Regarding the second question, we hypothesized that the higher the accuracy on the componential ability questions is on the first or second attempts, the higher is the growth in the standardized reading comprehension test. Larger effects were anticipated for first than second attempts.

2. Methods

2.1. Procedure

Dutch third to fifth graders took a computerized assessment with a dynamic approach in autumn 2018. This was conducted within three weeks during lecture times in five sessions. The sessions were scheduled and guided by teachers under researchers’ instructions. About three months prior and after the assessment, participants’ reading comprehension was assessed with a standardized test as part of the Dutch monitoring system.

2.2. Participants

In our study, 407 Dutch third to fifth graders participated (age: 8–11 years). They came from 22 classrooms across six schools in a suburban municipality in the East of the Netherlands. One school with two classes and three classrooms from two other schools (n = 69 participants) dropped out during the study and were excluded. To keep the same sample across analyses, we excluded participants without complete observations on the standardized reading comprehension tests (n = 25) or assessment (n = 42). The observations were missing at random (Little’s MCAR test: \( \chi^2 (26) = 28.76, p = .32 \)). We did not exclude children with language-related or developmental problems to reflect an authentic classroom in the Netherlands. The final sample consisted of 275 children (third graders n = 91, fourth graders n = 78, fifth graders n = 106) from 17 classrooms across five schools. The study sample performed slightly higher than the national sample on a standardized reading comprehension test of the Dutch monitoring system at the end of the former school year (\( t(5045) = -5.22, p = .01, d = 0.29 \)) and midterms of the current grade (\( t(8754) = -4.39, p = .01, d = 0.25 \)). The national sample was based on different participants at pretest (n = 7772) and posttest (n = 8481). The study sample’s growth in the reading comprehension test (\( t(274) = 11.10, p < .001, d = 0.46 \)) was comparable to the national sample’s growth (t

3
abilities across grades into account, a di-

texts, three assessment versions were developed with respectively 25
administered in third, fourth, and

2.3.1.1. Adapted assessment di-

the questions were answered. There was no time restriction.

For the orthographic knowledge, vocabulary, and sen-
tence-integration questions, feedback was

2.3.1.2. Question types. The assessment included four question types:

Orthographic knowledge questions. In order to measure the quality of
orthographic representations for three words from the text, the child
had to type each word after auditory presentation while the words
were covered in the text. If a word was spelled incorrectly, the
correct form was flashed for 3 s. After the correct form disappeared,
the child could make a second attempt. Cronbach’s alphas were 0.92
to 0.94, which indicated very high reliability.

Vocabulary questions. The quality of semantic representations for two
words from the text was measured via multiple-choice questions.

Fig. 2. Feedback loop for the orthographic knowledge, vocabulary, and sen-
tence-integration questions in the computerized assessment with a dynamic
approach.

Successively for each word, the child had to choose the correct de-

inition from three options. The word was printed in bold in the text.

If the response was incorrect, a picture was provided to resemble the
meaning before a new response could be given. The reliability of the

task was high (Cronbach’s alphas = 0.82 to 0.87).

Sentence-integration questions. To evaluate sentence-integration

abilities, the child had to answer one multiple-choice question. This
required the integration of explicitly provided information across
several phrases or sentences. There were four answer options. To
direct the reader’s attention, the text part which required integration
was highlighted. If the answer was incorrect, another part was
highlighted, which helped finding the correct answer. After in-
correct second attempts, a brief explanation of the correct answer
was shown. Cronbach’s alphas were between 0.78 and 0.88, which
displayed acceptable to high reliability.

Global text comprehension questions. The overall text comprehension
was measured with one multiple-choice question. The participants
had to select from four answer options the sentence which best re-

ected the main idea of the text or the best heading of the text. The
task proved to be sufficiently reliable with Cronbach’s alphas of 0.76
to 0.85.

More information about the development of the assessment can be
found in Den Ouden et al. (2019) and example questions are provided
in Supplement B.

2.3.1.3. Operationalization. For each participant, the response accuracy
on the first and second attempt after feedback for each question type
was operationalized in two ways.

2.3.1.3.1. Text level. We summarized the response accuracy on first
and second attempts for all items of each question type at the text level
because we could use only one score respectively for the prediction of
global text comprehension of a text. Each item was scored as 2 if it was
correct on the first attempt, as 1 if it was correct on the second attempt,
and as 0 if it was incorrect on the second attempt. Sum scores for all
items of a question type were built by either only considering the first
attempts (i.e., correct second attempts scored as 0) or both the first
and second attempts (i.e., correct second attempts scored as 1), as
exemplified in Table 1. A similar approach was taken by Elleman
et al. (2011).

Accuracy on the first attempts: Sum score for the first attempts

Table 1

Example for the sum score options for one text in the assessment with a dy-
namic approach.

<table>
<thead>
<tr>
<th>Scoring and sum score option</th>
<th>Performance on each question type for one text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Orthographic knowledge</td>
<td>Vocabulary</td>
</tr>
<tr>
<td>Item scoring for first attempts</td>
<td>Item 1 = 0</td>
</tr>
<tr>
<td>Item 2 = 0</td>
<td>Item 2 = 0</td>
</tr>
<tr>
<td>Item 3 = 2</td>
<td>Item 3 = 2</td>
</tr>
<tr>
<td>Item scoring for first and second attempts</td>
<td>Item 1 = 1</td>
</tr>
<tr>
<td>Item 2 = 0</td>
<td>Item 2 = 0</td>
</tr>
<tr>
<td>Item 3 = 2</td>
<td>Item 3 = 2</td>
</tr>
<tr>
<td>Sum score for first attempts</td>
<td>2</td>
</tr>
<tr>
<td>Sum score for first and second attempts</td>
<td>3</td>
</tr>
</tbody>
</table>

Note. Item scoring for first attempts (first attempt correct = 2, second attempt correct or incorrect = 0), item scoring for first and second attempts (first attempt correct = 2, second attempt correct = 1, second attempt incorrect = 0).
2.4. Data analysis

Tomesen et al. (2018). Accuracy on the first attempts: Percentage of correct responses on the first attempts
Accuracy after feedback: Percentage of correct responses from all attempts (i.e., questions which required feedback).

Due to the different assessment versions, the percentages of correct first attempts were not directly comparable between grades. However, in a pilot study, we found that the responses of third to fifth graders to each item from all assessment versions (item bank) could well be described by the one parameter logistic model (OPLM; Verhelst & Glas, 1995). With this model, it is possible to estimate the ability on the same scale with every possible subset of items from the item bank. The scores on the three assessment versions can thus be transformed into so-called ability scores which are comparable even if different assessment versions are being administered. The ability scores can further be transformed into so-called bank scores, which represent the expected percentage of correct responses of test-takers if they would have been assessed with all items from the item bank (Hambleton et al., 1991). In our study, we used these bank scores for the percentage of correct first attempts. The percentage of correct second attempts was based on the raw scores.

2.3.2. Standardized reading comprehension test

The Dutch primary school curriculum requires the regular assessment of reading comprehension. For this purpose, the Dutch National Institute for Educational Measurement (Cito) offers a student progress monitoring system with assessments for the midterm and end of a grade. As the assessments are based on item-response-theory calibrated item banks, the scores across grades are estimated on one ability scale.

The analyses were conducted with the package lme4 in R (Bates, Mächler, et al., 2015; R Core Team, 2018). To answer the first research question, we investigated with a generalized linear mixed effect model how the accuracy on the first attempts and after feedback for each componential ability question of a text affected the probability to answer the respective global text comprehension questions correctly. The accuracy on the first attempts and after feedback were highly correlated. Therefore, we used two separate models with either the sum scores for the first attempts only or both the first and second attempts per predictor. By adding random intercepts for participants and texts, we took the dependence of observations within participants (ICC = 0.19) and texts (ICC = 0.08) into account. The low dependence of observations within classes, schools, and grades (ICC ≤ 0.02) was not modeled. Random slopes and the random correlation terms of the experimental predictors were added if they significantly improved the model fit (p < .05; Barr et al., 2013). We tested the significance of fixed effects with the function mixed of the package afex (Singmann et al., 2018). This computes likelihood ratio tests between models with the predictor of interest and models without the predictor of interest. It is examined if the addition of a predictor significantly increases the model fit. Odd ratios of 1.68, 3.47, and 6.71 were considered respectively small, medium, and large effects (Chen et al., 2010).

To partial out the effect of the accuracy after feedback, we compared the fit of a model including the sum scores for the first and second attempts for all question types (baseline) to models from which respectively the second attempts for the question type of interest was excluded (i.e., sum score for only the first attempts as fixed effect), while keeping the models comparable otherwise. By this, the accuracy after feedback effect was investigated on top of accuracy on the first attempts for the same question type as well as the accuracy on the first attempts and after feedback for all other question types. An example is presented below:

Fixed predictors of the baseline model:
- sum score for the first and second attempts on the orthographic knowledge questions + sum score for the first and second attempts on the vocabulary questions + sum score for the first and second attempts on the sentence-integration questions

Fixed predictors of the comparison model for vocabulary questions:
- sum score for the first and second attempts on the orthographic knowledge question + sum score for the first attempts on the vocabulary question + sum score for the first and second attempts on the sentence-integration questions.

Since the models were not nested, likelihood ratio tests were not possible but goodness-of-fit statistics were compared. The goodness-of-fit is often described with the Akaike’s information criterion (AIC) and Bayesian information criterion (BIC), which indicate the amount of unexplained variance corrected for the number of predictors in the model (Burnham & Anderson, 2002; Field et al., 2012). Thus, the better fitting model has smaller AIC and BIC values compared to another model. BIC can be considered similar to the Bayes factor (albeit being on a different scale), which quantifies the evidence for an alternative hypothesis in contrast to the null hypothesis (Burnham & Anderson, 2002; Kass & Raftery, 1995). BIC differences can, therefore, be used to test hypotheses. A worse fit of the model without second attempts for the question type of interest (higher BIC) than the baseline model was considered positive evidence that correct second attempts after feedback for the question type under investigation increased the probability of a correct response to the global text comprehension question over and above all other effects. Model differences in BIC between 2 and 6 were considered positive evidence, between 6 and 10 strong positive evidence, and above 10 very strong positive evidence (Kass & Raftery, 1995).

To answer the second research question, we investigated with a linear mixed effect model how the posttest scores on the standardized reading comprehension test were predicted by the accuracy on the first attempts and after feedback for each componential ability question in the assessment over and above the pretest scores on the standardized reading comprehension test. In step 1 of the analyses, we added the reading comprehension scores at pretest as a predictor. In step 2, the accuracy on the first attempt and after feedback for each componential ability question were entered to the model. Four participants were excluded because they did not require feedback on some question types. The same procedure for the inclusion of random effects and significance testing was used as in the analysis for the first research question. Although the dependence of observations was relatively high in classes (ICC = 0.35) and schools (ICC = 0.19), we included only random intercepts for classes to remain model parsimony and convergence. Standardized estimates of 0.2, 0.3, and 0.5 were interpreted respectively as small, medium, and large effects (Field et al., 2012).
3. Results

3.1. Descriptive statistics

On average, one assessment session was finished by the participants in 22.47 min (SD = 16.44). The descriptive statistics for the accuracy on the first attempts and after feedback for each question type are provided in Table 2. The scores on the standardized reading comprehension test at pretest (M = 166.84, SD = 30.08) and posttest (M = 180.62, SD = 30.32) were strongly correlated with each other (r = 0.77) and with the percentage of correct responses to the global text comprehension questions (pretest: r = 0.66, posttest: r = 0.71). The reading comprehension test scores were strongly correlated with the percentage of correct first attempts for the componential ability questions (pretest: r = 0.58–0.63, posttest: r = 0.62–0.72) but weak to moderate with the performance after feedback (pretest: r = 0.24–0.40, posttest: r = 0.28–0.40). These correlations were respectively stronger for vocabulary and sentence-integration questions than orthographic knowledge questions. The percentage of correct responses to global text comprehension questions correlated strongly with the percentage of correct first attempts of other question types. This was the strongest for sentence-integration questions (r = 0.83), followed by vocabulary questions (r = 0.75), and the lowest for orthographic knowledge questions (r = 0.61). For the percentage of correct attempts after feedback, the correlations with global text comprehension were all moderate and slightly stronger for sentence-integration questions (r = 0.42) than for other question types (r = 0.35). All correlations were significant (p < .001). The inter-correlations were only weak when sum scores for texts were used due to the different scoring scale. A detailed overview is provided in Supplement C.

3.2. Prediction of global text comprehension by componential abilities within texts

The probability to correctly answer the global text comprehension question of a text was increased by each correctly answered orthographic knowledge, vocabulary, or sentence-integration question on the first attempt within the same text, as indicated by significant effects of the sum scores for the first attempts in Table 3. Thus, the higher the accuracy on the first attempts was, the higher was the probability of a correct response to the global text comprehension question. Although the increase of the probability by a single correctly answered component question can be considered small (OR = 1.07–1.31), this can sum up to large effects if all six questions were answered correctly on the first attempt.

In the second model, the sum scores for the first and second attempts were used as predictors. This unifies the accuracy on the first attempts and after feedback. A higher sum score for the first and second attempts on the vocabulary or sentence-integration questions led to an increased probability to answer the global text comprehension question correctly. The effect of orthographic knowledge questions was only marginal (p = .053). The difference between the models with and without the second attempts in Table 3 indicated the effect of the accuracy after feedback. For example, the increase in odd ratios was 0.15 by each correctly answered sentence-integration question on the second attempt. To quantify the evidence for an effect of the accuracy after feedback, the goodness-of-fit was compared between a baseline model with the second attempts in the sum scores for all question types and an identical model besides excluding the second attempts for the question type under investigation. The goodness-of-fit statistics AIC and BIC of each model is provided in Table 4. Including the second attempts on the sentence-integration questions led to a decrease of 8.50 compared to the model without second attempts. This can be considered strong evidence that a higher accuracy after feedback for the sentence-integration questions increased the probability to correctly answer the global text comprehension question. However, including second attempts for the orthographic knowledge or vocabulary questions did not lead to a decrease of the BIC larger or equal to 2. Thus, there was no evidence for an accuracy after feedback effect for these question types.

3.3. Prediction of growth in reading comprehension by componential abilities

There was a moderate autoregressive effect (b = 0.43) for the standardized reading comprehension test scores in Step 2, as presented in Table 5. On top of this, the percentage of correct first attempts for the orthographic knowledge, vocabulary and sentence-integration questions each explained unique variance in the posttest scores of the standardized reading comprehension test in Step 2 with small effect sizes (b = 0.10–0.20). The higher the accuracy was on the first attempts, the higher was the growth in the standardized reading comprehension test from pretest to posttest. For incorrectly answered decoding, vocabulary, and sentence-integration questions, the accuracy after feedback did not significantly influence the growth. The performance on the componential ability questions in the assessment overall explained 12% of the variance in the posttest scores on the standardized reading comprehension tests. Despite a strong correlation between the percentages of correct first attempts for the vocabulary and sentence-integration questions, all VIF values were below 3. This is not considered evidence for multicollinearity influences (Field et al., 2012; Hair et al., 1995).

Table 2

Means and standard deviations for each question type in the assessment with a dynamic approach.

<table>
<thead>
<tr>
<th>Question types</th>
<th>Percentage of correct responses</th>
<th>Sum scores for each text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First attempts</td>
<td>Second attempts after feedback</td>
</tr>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Orthographic knowledge</td>
<td>60.38</td>
<td>17.10</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>73.94</td>
<td>13.57</td>
</tr>
<tr>
<td>Sentence-integration</td>
<td>61.13</td>
<td>19.03</td>
</tr>
<tr>
<td>Global text comprehension</td>
<td>57.79</td>
<td>19.52</td>
</tr>
</tbody>
</table>

Note. n = 275 (n = 274 for the percentage of correct second attempts for the vocabulary questions, n = 272 for the percentage of correct second attempts for the sentence-integration questions). The percentage of correct first attempts was estimated on one scale for the different assessment versions with item response theory. The percentage of correct second attempts was based on the raw data and the number of all incorrect first attempts. For the sum scores, we summed up the item scores for each question type for the first attempts (0 = incorrect, 2 = correct) or first and second attempts (0 = incorrect after feedback, 1 = correct after feedback, 2 = correct without feedback). The possible score ranges were for the orthographic knowledge questions 0–6, vocabulary questions 0–4, sentence-integration and global text comprehension questions 0–2. n.a. = not available.

For all analyses, we centered continuous predictors. If necessary, bound optimization by quadratic approximation was used. If not reported differently, model assumptions were met (Field et al., 2012; Hartig, 2018; Nieuwenhuis, Te Grotenhuis, & Pelzer, 2012).
Note. The models without the second attempt in the sum scores on a particular question type were otherwise identical to the baseline model. $AIC = \text{Akaike's information criterion, } BIC = \text{Bayesian information criterion.}$
standardized reading comprehension test on top of its autoregressive effect. Thus, each of the three components does not only explain variance in comprehension within the same texts but also in the development on general reading comprehension ability. In the latter, we tested the contribution of each component in quite a strict way because the performance on the standardized reading comprehension test at pretest may cover a lot of the variance in the componential abilities (Richter et al., 2002; McMaster et al., 2012). The assessment seemed to tap into additional abilities which were not covered by the standardized reading comprehension test.

4.2. Prediction of variance in reading comprehension by the accuracy after feedback

Over and above the first attempts, we found an effect of the accuracy after feedback on global text comprehension for the sentence-integration questions but not for orthographic knowledge and vocabulary questions. This means that only the accuracy after feedback for sentence-integration questions could differentiate between children of various abilities in global text comprehension. It can be concluded that more fine-grained profiling of sentence-integration abilities was possible. Children who could answer the questions correctly after the relevant text part was highlighted did not have difficulties with the integration itself but with identifying the proper information. In contrast, children who did not answer correctly after feedback had problems with integrating the information even after being directed to it. This may reflect a qualitative difference, which may ask for a different instructional focus. Similar subgroups were also reported previously (Cain et al., 2001; McMaster et al., 2012).

It is unclear why the accuracy after feedback on the orthographic knowledge and vocabulary questions did not differentiate between children’s global text comprehension. Similar non-significant findings were also reported by Den Ouden et al. (2019), although they considered aggregated scores across texts instead of within texts. One possible explanation is that the word-level feedback was not sufficiently integrated to the text. The pictures provided after incorrect vocabulary questions were not matched with the text context (Carney & Levin, 2002), and the feedback after incorrect orthographic knowledge questions just focused on the form. Processing the word-level feedback, thus, required the reader to shift the focus away from the text, whereas highlighting sentences was directly linked to the text (Sweller, 2011).

Other reasons might lie in the nature of feedback (Hattie & Timperley, 2007; Shute, 2008). While the correct answer was previewed briefly for the orthographic knowledge questions, compensation was offered in the vocabulary and sentence-integration questions (i.e., bypassing reading via pictures, taking over the information-selection step). This may cause dissociations in what the accuracy after feedback measured and how this mattered for global text comprehension. While sentence-integration feedback helps overcoming an obstacle in the comprehension process, the accuracy for word-level tasks after feedback may rather evaluate a lexical-knowledge level. The word-level feedback was adapted from training studies and might still be useful to differentiate reading comprehension abilities, feedback should less differentiate in what the accuracy after feedback measured and how this mattered for global text comprehension. While sentence-integration feedback helps overcoming an obstacle in the comprehension process, the accuracy for word-level tasks after feedback may rather evaluate a lexical-knowledge level. The word-level feedback was adapted from training studies and might still be useful to differentiate reading comprehension abilities, feedback should less differentiate in what the accuracy after feedback measured and how this mattered for global text comprehension.

Table 5
Prediction of standardized reading comprehension test scores at posttest.

<table>
<thead>
<tr>
<th>Predictors</th>
<th>( \chi^2 )</th>
<th>b</th>
<th>B (SE)</th>
<th>95% CI for B</th>
<th>( \chi^2 ) (df) model comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Step 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>182.68 (2.56)</td>
<td>0.72</td>
<td>0.73 (0.07)</td>
<td>0.59</td>
<td>0.87</td>
</tr>
<tr>
<td>Pretest reading comprehension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random for classrooms:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>83.91</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest reading comprehension</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residuals</td>
<td>287.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Step 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>181.97 (2.04)</td>
<td>0.43</td>
<td>0.43 (0.06)</td>
<td>0.31</td>
<td>0.56</td>
</tr>
<tr>
<td>Pretest reading comprehension</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Orthographic knowledge first attempt</td>
<td>0.10</td>
<td></td>
<td>0.18 (0.08)</td>
<td>0.02</td>
<td>0.33</td>
</tr>
<tr>
<td>Orthographic knowledge after feedback</td>
<td>-0.05</td>
<td></td>
<td>-0.11 (0.08)</td>
<td>-0.28</td>
<td>0.06</td>
</tr>
<tr>
<td>Vocabulary after feedback</td>
<td>0.19</td>
<td></td>
<td>0.43 (0.14)</td>
<td>0.16</td>
<td>0.69</td>
</tr>
<tr>
<td>Sentence-integration first attempt</td>
<td>-0.02</td>
<td></td>
<td>-0.05 (0.07)</td>
<td>-0.19</td>
<td>0.10</td>
</tr>
<tr>
<td>Sentence-integration after feedback</td>
<td>0.20</td>
<td></td>
<td>0.32 (0.09)</td>
<td>0.15</td>
<td>0.49</td>
</tr>
<tr>
<td>Random for classrooms:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>50.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pretest reading comprehension</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residuals</td>
<td>222.85</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. n = 271 (Four participants were excluded because they did not require feedback on some question types). Step 1: \( R^2_m = 0.53, R^2 = 0.68 \); Step 2: \( R^2_m = 0.65, R^2 = 0.74 \). \( s^2 \) = variance, \( b \) = standardized estimate, \( B \) = unstandardized estimate. CI = confidence interval.

- * \( p < .05 \)
- ** \( p < .01 \)
- *** \( p < .001 \)
children's global text comprehension within the same texts, it is not surprising that this effect was not found for the comprehension of other texts in the standardized test over time. This may also explain why the above-described accuracy after feedback effect for sentence-integration questions did not remain significant. The effect was, probably, too small. Also, in the study of Elleman et al. (2011), the performance on sentence-integration questions including the responsiveness to feedback explained only a small amount of variance in a standardized reading comprehension test. This might have been significant because they did not control for the accuracy on the first attempts in contrast to our study.

4.3. Limitations and future investigations

This study has some limitations due to the type of data the assessment entailed. The first and second attempts for different question types of a text were not independent. Therefore, we used random person and text effects in our analyses. For the prediction of growth in the standardized reading comprehension test, however, we had to summarize the performance for each question type because there was only one score available as dependent variable. Model comparisons of goodness-of-fits had to be conducted to circumvent the correlations between first and second attempts at test level. The models were not identical in their underlying covariance structure but differed with respect to coding of one parameter (sum score with or without second attempt). Since the range of the scores with or without the second attempts was comparable, this is a very small difference.

Despite observing a large variety of different profiles in our study (see Supplement D), we could not provide empirical evidence for them. As about two-thirds of the participants did not perform lower than $-1$ SD from the grade level mean on any question type, the remaining sample was considered too low to evidence with a cluster analysis more fine-grained profiles than those commensurate with the simple view of reading. Follow-up research should include more same-aged children with reading comprehension problems or should increase the assessment sensitivity toward differences between higher-skilled children. It is also not clear if the componential measures within the same text can indeed capture the interactive online-processes in reading comprehension better than isolated tests. In future, the profiles identified with our dynamic approach or with isolated tests could be compared with respect to congruency and effectiveness of profile-adapted instructions. Other feedback forms as well as its influence on response behavior may be considered to better understand the role of feedback for profiling and to individualize the assessment (Van der Linden & Glas, 2002).

5. Conclusion and implications

We found evidence for the suitability of a dynamic approach for profiling of children's instructional needs in reading comprehension in terms of the focus and intensity of instruction. The accuracy for orthographic knowledge, vocabulary, and sentence-integration questions on the first attempt predicted the global text comprehension within texts and the growth in a standardized reading comprehension test. The individual performance pattern across the three componential abilities can provide an insight into the required focus of instruction in reading comprehension. Assessing the componential abilities within the same text seemed to tap into additional skills, which are not captured by a standardized reading comprehension test. Moreover, the accuracy after feedback for sentence-integration questions further differentiated children's reading comprehension abilities. This difference seemed to indicate a different instructional focus in sentence-integration (i.e., identifying vs. integrating information). Although the accuracy after feedback on the orthographic knowledge and vocabulary question did not differentiate children's reading comprehension abilities, it may still inform on the required intensity of word-level instructions. More process-related feedback may contribute to profiling of reading comprehension.

As the presented assessment takes a rather restrictive view on what entails reading comprehension at text level (i.e., identifying the main message), it may not replace the tasks of existent standardized tests. It should rather be seen as a supplemental tool to better profile the individual reader compared to standardized and isolated tests. Although the increasing role of assessments in children's education is criticizable, the assessment can contribute to the efficiency of instructions at schools.

Author note

Sophie Gruhn, Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands; Eliane Segers, Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands, Instructional Technology, University of Twente, Enschede, The Netherlands; Jos Keuning, Cito, Institute for Educational Measurement, Arnhem, The Netherlands; Ludo Verhoeven, Behavioural Science Institute, Radboud University, Nijmegen, The Netherlands, Faculty of Arts, University of Curacao, Curacao.

Funding

This work was supported by The Netherlands Initiative for Education Research (National Regieorgaan Onderwijsonderzoek) [grant number NWO 405-15-548].

Declaration of competing interest

None.

Acknowledgements

This work was based on a cooperation between the Radboud University (Ludo Verhoeven, Eliane Segers, Sophie Gruhn), Cito and Twente University (Theo Eggen, Jos Keuning, Marije den Ouden), Expertisecentrum Nederlands (Nicole Heister-Swart), and Kennisinstituut voor Taalontwikkeling (ITTA, Femke Scheltinga).

Data accessibility statement

Data available on request from the authors.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.lindif.2020.101923.

References

S. Gruhn, et al. Learning and Individual Differences 82 (2020) 101923


