Quality Assurance in the National Tests of English: Investigating What Makes Reading Difficult

Angela Hasselgreen, Craig Grocott and Torbjørn Torsheim

This article presents two studies, the second building on the findings of the first, which investigate the relationship between features of test items/texts used in the NTE (National Tests of English) test of reading and the difficulty of items. It will look for indications that the features found to influence difficulty are reflected in the progression of descriptors of reading ability in the Mastery levels used in reporting test results.

1. Background

The NTE is taken by virtually all pupils in Norwegian schools, early in 5th and 8th grades. The tests, which are delivered online, consist of 40 to 50 items, covering a wide range of difficulty and a variety of formats. The tests are designed to test general reading in English. The University of Bergen, in association with Uni Research, has been responsible for the development, trialling and analysis of the NTE. The authors have been involved in all of these processes.

According to the Norwegian Directorate for Education and Training (*Utdanningsdirektoratet)*:

> The purpose of the national tests is to provide schools with information on pupils' basic skills in […] English. Information from the tests should provide a basis for ongoing assessment and quality development at all levels of the education system" (translation) (Nasjonale prøver 2016)

Remaining Relevant: Modern Language Studies Today

Bergen Language and Linguistics Studies · vol. 7 · 2017 DOI: http://dx.doi.org/10.15845/bells.v7i0

Teachers receive advice in the test guidelines on how to use the individual pupil's test results to support his/her development in English reading. The test scores are converted to a series of mastery levels, three for 5th grade and five for 8th grade (see Appendix). These contain detailed descriptors of which subskills a pupil at the particular level can be expected to master (Mestringsbeskrivelser 2016).

The team responsible for developing the test, including the guidelines and level descriptors, have recently undertaken a series of studies on test item data. The purpose of these studies was, primarily, to establish what pupils had to 'do' (or which subskills were necessary) in order to correctly answer specific test items, and to determine how these subskills are associated with item difficulty. This in turn should inform us both in the creation of items intended for specific levels of difficulty and, importantly, in adjusting the descriptors for the mastery levels

Before embarking on the studies themselves, investigating the link between features and item difficulty in the NTE, we will first consider what other researchers have concluded with regard to the features that make reading texts and tasks difficult.

2. Features of texts/tasks found to be associated with difficulty

Difficulty in reading can be caused by features inherent to the text itself, by properties of the task the reader is required to carry out, or to both.[1]

2.1. Features of texts

Studies such as Alderson (2000) and Crossley, Greenfield & McNamara (2008) focus on features of the overall text which have been found to contribute to difficulty in reading.

Alderson (2000) identifies a number of features of a text that affect difficulty. The first is vocabulary difficulty, which "has consistently been shown to have an effect on understanding for first-language readers as well as second-language readers" (Alderson 2000, 69). However, the author notes the effect of topic (un)familiarity on vocabulary difficulty. The second feature identified is syntactic complexity; i.e. the "opacity and heaviness of the constituent structures which make it difficult for readers to parse syntax" (Alderson 2000, 69). However, the author notes that

1 Here 'item' is used to refer to a 'text plus task', while 'task' refers only to what the pupil has to 'do' and not the text itself.

"simplifying syntax does not necessarily make texts more readable, since a thorough syntactic analysis of text may be unnecessary" (Alderson 2000, 73). Topic, cohesion, coherence and 'readability' were also cited, with a warning of simplistic assumptions, due to interactions between features.

Crossley, Greenfield & McNamara (2008) found that a high count on three variables was likely to indicate an increase in the ease of reading, by reducing the cognitive load for the reader. The first variable is the lexical index, based on the proportion of high-frequency words. The second variable is the syntactic index, based on similarity of the syntactic constructions in the sentences in the text. The third and final variable is the meaning-construction index, based on how often content words overlapped in adjacent sentences.

2.2. Features of tasks

In contrast to predictors of reading difficulty inherent to overall *texts*, other studies have investigated features that predicated difficulty when carrying out particular *tasks* in reading tests. Lumley et al's 2012 study focused on the parts of a text that the reader needed to understand in order to do the task, and found that five variables correlated significantly with item difficulty: competing information, relationship between the task and required information, concreteness of information, familiarity of information, and reference to information from outside the text (world knowledge, personal beliefs, ideas and opinions).

In another study, Gao and Rogers (2011) concluded that difficulty in carrying out reading test tasks is increased by several features: the number of plausible distractors, the requirement of high-level inferencing, the degree of syntactic knowledge required, the need to use contextual clues to work out the meaning of unknown vocabulary, and the degree to which the required information is spread throughout the text.

This wide range of predictors of difficulty associated with tasks can perhaps be reduced to two basic types of features. The first are features associated with *the information required by the task*—its location, its concreteness or familiarity and the relationship or degree of overlap with the wording of the task. The second are features associated with *the level of reading processing required* to do the task—recognising and understanding lexis, using syntax (and morphology) to understand sentences, con-

structing meaning across texts, and using outside information or intuition to make inferences. This second set of features is reflected in Khalifa and Weir's (2009: 43) hierarchy of levels of processing, as follows:

Constructing a representation across texts
Constructing a representation of the whole text
Inferencing
Creating propositional meaning at clause and sentence level
Syntactic parsing
Accessing lexical meaning
Word recognition

3. Study one

The first study was carried out using data from 5th-grade tests, with the aim of identifying features of items which appear to influence difficulty. A sample of items whose p-values were known formed the basis of the study, and the features studied included a number of those identified above as being associated with difficulty. Where subjective judgements were necessary, the items were rated by seven trained raters and an analysis was carried out to investigate the associations of item features and p-values.

3.1. Sample

The study was based on 176 test items. As the number of variables/predictors investigated was to be 17, a set of 176 items was considered an appropriate size, in the light of the statement by Crossley et al (2008) that: "Generally, a minimum of 10 cases of data for each predictor is considered to be accurate" (2008, 482). It was considered preferable to use items used in the past two-three years in order to best reflect current items. As each actual test consists of no more than 50 items, the items for the study were chosen not only from test data but also from piloting data (i.e. items whose p-values were known but which had not been used so far in tests). The items were chosen with the aim of putting together a set which reflected the distribution of format/task types typically used in a test.

The 17 potential variables included eight that required rater judgements. These eight included four which were related to language/reading processing—*recognising vocabulary, understanding simple sentences, un-*

derstanding complex sentences, and *making links between sentences.* The remaining four were more explicitly associated with the task—*reading to find information/detail*, *reading to grasp overall meaning/main point* and *inferencing*, as well as the presence of *competing information*.

The nine categories which did not require rater judgements, being inherent features of the tasks/texts, included *text length* as well as eight categories of *format/task type*, for example, *move object*, or *who could say*.

3.2. Procedure

The training of seven raters was carried out over four sessions, with individual rating followed by discussion. The description of predictor categories was adjusted after each round, and in some cases the categories themselves were changed. The final rating was carried out by all seven raters individually. The raters were issued with a spreadsheet with the instruction to place a cross under all of the features that they felt were present in the individual items, without a limit as to how many crosses could be placed per text. If the rater chose the category 'sentence', the category 'vocabulary' would also be chosen, as one cannot understand a sentence without understanding the vocabulary within it.

3.3. Analysis

Inter-rater correlations were calculated for all item features, using intraclass correlations from ANOVA.

As tasks might tap several processing components, we computed task-feature combinations. Based on the study classification scheme, each item was classified according to the highest level of complexity. The low-level complexity group consisted of items requiring vocabulary only. The intermediate-level group consisted of items requiring sentence-clause combinations, reading to form overall meaning, and linking sentences. The high-level group consisted of items requiring inferencing and including competing information. The association between level of complexity and p-correct was tested using Oneway ANOVA, as was the overall effect of task complexity.

To identify the relative importance of specific task features, a multiple regression analysis was used, regressing p-values on all features in order to establish which items significantly contributed to the prediction of p-correct, and hence difficulty.

3.4. Results

The statistical analysis showed acceptable inter-rater correlations (over .70) on all categories, using intra-class correlations from ANOVA, which suggests that the mean of the ratings was reliable.

The Oneway ANOVA of the task complexity groups showed that task complexity had a statistically significant effect, $F(2, 173) = 19.21$, $p < 0.001$, eta-squared= 0.18, indicating that the p-correct varied as a function of task-complexity group. The mean p-correct was lowest in the high-complexity group ($M = 0.43$), and highest in the low-complexity group ($M = 0.65$).

To assess the specific importance of all task features, a multiple regression model was tested using p-correct as the dependent variable and task features as independent dummy variables. Vocabulary was used as the reference category. The results of the multiple regression analysis are displayed in table 1. The following processes were established as having a positive effect on difficulty, using multiple regression: the requirement to understand a *sentence/clause* (involving syntax, morphology and function words and reflected in M-2 descriptors), and the requirement to make *inferences* (reflected in M-3 descriptors).

A separate analysis of item characteristics revealed that text length was a very strong predictor of difficulty for items overall. However, this was not the case in items requiring vocabulary understanding alone. Another characteristic that was a good predictor or difficulty was task format; those based on pictures were less difficult than those based on text alone.

Table 1. Model Summary from Multiple Regression of p-correct Regressed on Task Features

Variable	B	SE	Stardardized Beta
Intercept	0.65	0.026	-
Sentence/clause	-0.17	0.033	-0.38***
Complex sentence	-0.02	0.036	-0.03
Link sentences	0.05	0.043	0.09
Read for overall meaning/main pt	-0.14	0.105	-0.09
Inference	-0.16	0.061	-0.18**
Competing info	-0.06	0.035	-0.12

Note.** $p < 0.01$; * $p < 0.05$

3.5. Discussion

Limitations on the study were related to the fact that there were very few items testing overall meaning, and that the item data was taken from a number of sources (both real tests and pilot versions), which can affect the parity of item behaviour. The p-value of an item may depend on whether the item was used in testing or piloting; this can be attributed to students taking actual tests more seriously than pilot tests.

The findings of this study, to a large extent, reflect the discussion above regarding features that have been found by other researchers to predict difficulty. These include levels of cognitive processing as well as features of the task—the concreteness of information (Lumley et al, 2012), being associated here with the presence of pictures in the tasks.

A consequence of the study was that the test development team's awareness of what items appeared to measure increased. This led to a revision of the coding of items in the item bank and an adjustment to the information provided in the teacher guidelines regarding what items measure. Up to this point, this information had been given on a one-dimensional scale, with elements such as *understand vocabulary, link sentences, find main point*, or *find detail*. Each item had been coded as measuring one or more of these elements. This meant that several labels could be assigned to the same item; moreover, there was no limit as to how many categories an item could be assigned to, although some of them are mutually exclusive. There was an apparent need to separate the categories, which was undertaken in the following study, with a view to developing a clearer structure.

4. Study two

After completing the first study, which used data obtained from the 5th-grade items, this study was replicated using data from 8th-grade items, with some extended (or revised) elements.

The deeper understanding which resulted from the 5th-grade study, and the apparent need for separation of categories, led to a new, two-dimensional model for coding items, with levels of cognitive reading processing as one dimension, and the operation required by the task as the other. For each category, only one feature could be chosen per item.

Dimension 1- the levels of reading processing are presented as a hierarchy, with each level building on the one preceding it, as follows:

1. VOCABULARY: *Understand vocabulary-* understand a word or phrase, possibly with the support of the context
2. SENTENCE: *Understand sentence(s)-* understand a sentence/clause, or a number of adjacent sentences/clauses
3. LINK: *Link sentences/parts of the text-* make the connection between sentences which are separated in the text. This can also involve linking between different types of text, e.g. diagrams.

Given that the categories assigned to Dimension 1 are hierarchical in nature, they cannot be mutually exclusive in the way that the categories in the second dimension are. Nevertheless, each item is placed in only one category, since one can assume that, per definition, the categories higher up in the hierarchy incorporate the 'lower' categories. For example, one cannot link sentences which are separated in a text without understanding the sentences themselves. There can be exceptions to this rule in theory, but these are rare enough to accept this hierarchical structure. Given that no texts were very long (maximum, ca 350 words) and no items involved linking information from several texts, these categories were comprehensive in that they covered all items.

Dimension 2—the operation required by the task is expressed as five mutually exclusive categories:

1. INFO/DETAIL: *Find (specific) information/understand (specific) detail-* find a specific piece of information or detail which is given in a text or picture.
2. MAIN POINT: *Understand the main point-* identify the main point of a text or a section of a text.
3. INTERPRET: *Interpret and understand-* interpret or have a more intuitive understanding of the text—the information required is not to be found directly in the text.
4. GRAMMAR: *Understand/use grammatical structures-* select or provide a particular grammatical structure (syntax/morphology/function word).
5. COHESION: *Understand cohesion*—put a series of disconnected paragraphs in a text into the right order.

The categories in Dimension 2 are considered to be mutually exclusive because they represent distinct item types and do not necessarily incorporate one another in a hierarchical structure as with those features in the first category. Some of the operations required by the task do, however, have an intrinsic connection to certain reading processing levels in Dimension 1. For example, if an item has *main point* as its key operation, the level of reading processing would automatically be classed as *link*, since different parts of the text have to be understood in order to ascertain what the main point of the text is. The only feature in the second dimension that could be argued not to be exclusive is the first, *info/detail*, as one must understand individual details in order to, for example, interpret the main point of a text. However, it is necessary to include the *info/detail* feature as a separate 'operation' category, in order to cover items that require *only* this operation.

The categories *grammar* and *cohesion* are unique in that they are both tied to specific task types. Grammar, as explained above, usually involves filling a gap in a sentence, either with given alternatives or without, in order to demonstrate the candidate's understanding of specific grammatical structures. It is not tied to one specific level of reading processing, but is virtually always paired with *understand* sentence; it would be unlikely to be paired with *link*. *Cohesion* on the other hand is always tied to *link*, as the candidate must be able to connect the different parts of an entire (usually longer) text.

Study two was carried out in the autumn of 2016, after the new coding of items had been established. This study differed from the earlier one in that the 8th grade test was the object, rather than the 5th-grade test. This was largely because the items offered more scope in the range of what was tested; it was more common at this level to test for overall meaning, inferencing and cohesion. In addition, the 8th-grade tests spanned all five mastery levels. Data was provided showing the calculated mastery level assigned to each item.

4.1. Sample

It was decided that the most logical and practical variables to investigate should include the eight categories represented in our two-dimensional model. This would make the findings directly relevant, as these categories are those referred to in the guidelines for teachers using the tests.

Moreover, as the team were familiar with coding items in these categories, this was expected to enhance rater reliability. As all eight categories involved features relating to actual tasks, rather than to the texts on which they are based, it was decided that one category relating to the entire text should be included in the dataset; therefore the *Flesch-Kincaid Grade Level Readability* index was included for each item. The reasoning behind including this was that it takes into account aspects of the text that are not necessarily reflected in the two dimensions explained earlier, such as sentence length and average word length (Flesch-Kincaid n.d.). These can be regarded as indirect measures of sentence and vocabulary complexity. Text length was also noted and included in the analysis since study one indicated that text length was a strong indicator of item difficulty.

As ten potential variables were to be investigated, it was not necessary to use data from more than roughly 100 items, following the principle of 10 raters per category described above. This meant that the actual data from two entire tests could be used. This had the advantage of using more comparable p-values, as piloting p-values tend to differ slightly from test p-values. In fact, the number of items used was 92, this being the number of test items in the 2014 and 2015 tests, on which the study was based.

In addition to p-values, data was also available for the mastery (M)-levels each item was assigned to. The M-level for each item had been established using item-difficulty estimates based on IRT calculations, combined with the percentage distribution of items across levels. The assignment of items to these levels was the responsibility of *Utdanningsdirektoratet*.

4.2. Procedure

All items were rated individually, going through the test online, with no access to the original coding. In deciding the coding on the first dimension, the raters had to consider what was strictly necessary, in terms of the level of processing, in order to do this task. If, for example, recognising a single word or phrase was sufficient to arrive at the answer, then the item was coded for *vocabulary*, regardless of how long or complex the text might be. For the operation dimension, the wording of the task, and how this related to the text, was crucial to the coding.

The rating was done in two stages. For a sample of approximately one quarter of the items, agreement was reached by consensus following the

individual rating. For the remaining items, individual ratings were registered and the final coding required that at least four of the seven raters were in agreement. On the occasional instances where this produced no clear result, a verdict was reached after discussion with the raters.

4.3. Analysis

A range of statistical methods were used in the analysis. Bivariate correlation was employed for all individual features to find associations between item features and difficulty. The relative importance of task dimensions was assessed through blockwise hierarchical regression, placing levels of reading processing in the first block and task operations in the second block. Finally, to detect possible non-additive effects, regression trees were estimated using item difficulty as the dependent variable, and all task features as predictive variables.

4.4. Results

When considering all ten features investigated, we estimated the association between item features and p-correct (item difficulty). The associations are shown in Table 2.

Table 2: Bivariate associations between task features and task difficulty (p correct)

Task feature	p-correct
Vocabulary	.185
Sentence	-0.174
Link	0.021
Info/detail	**.252***
Main point	0.45
Interpret	0.32
Grammar	**-0.380****
Cohesion	.006
Words	**.235***
Readability	**-.463****

Note.** $p < 0.01$; * $p < 0.05$

The task operations (Dimension 2) *understanding information/detail* (easy) and *grammar* (difficult) were significantly related to item difficulty. Features of the text—*length* and *readability* (roughly associated with

sentence and vocabulary complexity)—were also significantly related to item difficulty.

To understand the relative importance of the two dimensions of features, these were entered in a blockwise hierarchical regression analysis (not reported in table). Levels of reading processing were placed in the first block, whereas task operations were placed in the second block. The first block did not have any predictive value, as indicated by the nonsignificant change in R-squared. The task operations in the second block had a statistically significant increase in R-squared of 0.113 (*F-change*(3, 82)= 3.69, p< 0.015)). Adjusting for other variables, *Grammar* achieved statistical significance, with a standardised coefficient (Beta) of -0.36 (B =-0.209, t=3.23, p<0.002).

When considering only the task operation (Dimension 2), items were divided into three groups of hypothesized task complexity: items requiring finding *information/detail* as the least complex, items requiring *main point* and *interpret/inference* features as being of intermediate complexity, and items requiring *grammar* and *cohesion* as the highest level of complexity. According to this classification, 55 items belonged to the low-complexity group, 15 to the intermediate-complexity group, and 21 items to the high-complexity group.

A one-way ANOVA, using complexity classification as the independent factor predicting item difficulty (p-correct), revealed the significant main effect of feature complexity on item difficulty. It indicated that an increase in the complexity of items, in terms of operations, was associated with increased item difficulty. Thus, tasks requiring *information/detail* were easier than those in the *main point- interpret/inference* group, which in turn were easier than those in the *grammar–cohesion* group.

To examine the non-additive effects of item features, we used classification and regression trees to predict item difficulty. As the current sample size of items was quite low, only a learning sample was used. Thus, the replicability of the tree was not assessed.

The regression tree showed that the *grammar* items have the highest degree of difficulty, with indications that *main point* and *link* were also associated with high difficulty; *information/detail* items were shown to be most closely associated with low difficulty.

When considering the regression tree for p-correct, it could be seen clearly that items requiring the cognitive level *vocabulary* only were

significantly easier than those requiring understanding of *sentences*. Interestingly, where sentence understanding was necessary, items with the operation *main point* were shown to have a high level of difficulty.

To sum up, the results indicate that, with respect to reading processing (Dimension 1), the items requiring an understanding of *vocabulary* were clearly easier than those requiring an understanding of *sentences*. There was some indication that making *links* between sentences adds to difficulty. The main effect of reading processing was not strong. However, the regression tree for mastery level indicated some interactive effects: making links added to difficulty when the task did not involve grammar.

In terms of task operations (Dimension 2), items requiring finding *information/detail* emerged as the easiest, while those requiring finding the *main point* and *interpreting/inferencing* constitute a group of intermediate complexity, and *grammar* and *cohesion* items comprise the most complex group.

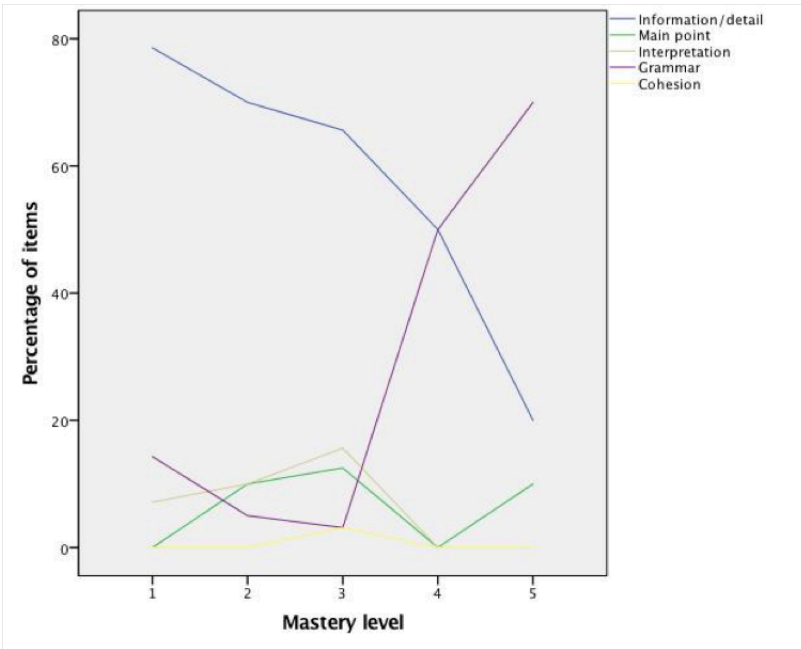

Figure 1. A visual representation of the distribution of item task features at different levels of mastery

It also emerged that the features of the text investigated—text length and readability—correlated significantly with text difficulty.

Whilst most of the analysis in study 2 refers to p-correct, further analysis indicated that the results on mastery levels were almost identical to the results reported for p-correct (such as correlation, hierarchical regression, and regression trees). The correlation between mastery level and p-correct is very strong, a negative correlation at around -0.91.

Figure 1 shows a visual representation of the distribution of item task features at different levels of mastery.

5. Implications of the findings of the studies

While it was beyond the scope of the studies carried out to investigate all the subskills of reading included in the mastery levels (Appendix), the studies have shown that, in the case of the features investigated, the difficulty associated with features is largely reflected in the descriptors for the M-levels.

In the case of reading processing (Dimension 1), the findings from a previous study were borne out to a large extent by studies one and two. Items requiring only an understanding of words and phrases were found to be easier than those requiring an understanding of a sentence or clause; the difficulty increases further when it is necessary to make links between non-adjacent sentences in a text. This finding is reflected in the mastery levels.

Vocabulary is, of course, essential to understanding any text, and is therefore mentioned at all M-levels. The nature of the vocabulary differs with the level, however, being most concrete and familiar at level 1, and steadily increasing in complexity throughout the levels. While the nature of the vocabulary was not included in our studies, this feature is commonly cited as a predictor of difficulty by other researchers, such as Alderson (2001). The understanding of sentences is not a requirement before level 2 where only simple sentences are mentioned, while more complex sentences are added at level 3. Again, this is a commonly cited feature of text difficulty, for example by Gao and Rogers (2011), but it should be noted that in study 2, where complex sentences were included as a feature, these were not found to significantly add to the difficulty. Linking information in different parts of a text is included at level 3 and is more firmly established at level 4.

Regarding Dimension 2, the operation required by the task, it was found that the simplest of these, *finding information/detail*, is clearly present as early as mastery level 2. Operations of intermediate difficulty, *understanding the main point* and *interpreting/inferencing*, emerge gradually at level 2–3 and 3–5 respectively. The two most difficult operations were found to be *cohesion* (ordering paragraphs) and, at the extreme end, *grammar* (involving choosing a grammatical form). The ability to relate paragraphs is mentioned in the level 3 descriptors, while making grammatical choices is not included until levels 4–5.

6. Further Study

Further study is needed to investigate the impact on difficulty of other features cited in the research, such as competing information. In addition, the fact that there were indications of a relationship between characteristics of the text itself (as opposed to features of the task) and task difficulty suggests that further investigation into this relationship would be useful. It would be of great interest and importance to establish whether or not characteristics of the text itself have less or as much (or even more) impact on the difficulty of the item than the features of tasks discussed in this paper.

Nevertheless, it can be concluded that, in the case of the features studied here, there is an indication that the way these affect difficulty is represented in the descriptors for the 5 mastery levels used in the National tests of English.

References

Alderson, J.C. 2000. *Assessing Reading*. Cambridge: Cambridge University Press.

Crossley, S.A., J. Greenfield, and D. McNamara. 2008. "Assessing Text Readability Using Cognitively Based Indices." *TESOL Quarterly* 42 (3):475–93.

Flesch-Kincaid n.d. "The Flesch-Kin Grade Level Readability Index." http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php (accessed 4 January, 2017).

Gao, L. and W.T. Rogers. 2011. "Use of Tree-based Regression in the Analysis of L2 Reading Test Items". *Language Testing* 28 (1):77–104.

Mestringsbeskrivelser 2016. "Mestringsbeskrivelser og hva nasjonale prøver måler," *Utdanningsdirektoratet*, (Norwegian Directorate for Education and Training), https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/mestringsbeskrivelser-og-hva-provene-maler/ (accessed 4 January, 2017).

Nasjonale prøver 2016. "Kva er nasjonale prøvar?" *Utdanningsdirektoratet*, (Norwegian Directorate for Education and Training), https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/om-nasjonale-prover/ (accessed January 4, 2017).

Plakans, L. and Z. Bilji. 2016. "Cohesion Features in ESL Reading: comparing beginning, intermediate and advanced textbooks". *Reading in a Foreign Languag,* 28 (1):79–100

Kahlifa , H. and C.J. Weir. 2009. *Examining Reading: Research and Practice in Assessing Second Language Reading. Studies in Language Testing 29.* Cambridge: Cambridge University Press.

Hasselgreen, A. and G. Caudwell. 2016. *Assessing the Language of Young Learners (*British Council Monographs 1). Sheffield: Equinox.

Lumley, T., A. Routitsky, J. Mendelovits and D. Ramalingam. 2012. "A Framework for Predicting Item Difficulty in Reading Tests." Paper presented at the American Educational Research Association Meeting, Vancouver, April 13–17.

Appendix—Mastery levels
(Mastery levels 1–3 apply to 5th grade tests. Mastery levels 1–5 apply to 8th grade tests.)

Mastery level 1
Pupils

Can understand some concrete, common words and expressions
Can find common, concrete words in a text
Can follow clear, simple instructions
Can link common, concrete words to pictures
Can make links between familiar, concrete words within a theme, e.g. *fish* and *acquarium*

Can recognise some learnt grammatical expressions and simple function words in context, e.g. personal pronouns.

Mastery level 2
Pupils

Can understand a number of common words and expressions
Can understand simple sentences
Can link simple sentences to pictures
Can make links between common words in a text, when they are within a theme
Can find specific details in a longer text
Can find simple synonyms in a short text
Can understand the main point In a simple text
Can find simple information even when there is some competing information in a text
Can navigate back and forth in a text to find information
Can draw simple conclusions when there is a good deal of support in the text
Can recognise and use some simple function words and grammatical structures in context

Mastery level 3
Pupils

Can understand rather abstract and less common words and expressions
Can construct meaning from some complex sentences
Can construct meaning from shorter and longer texts
Can understand the main point In a text
Can find information even when there is competing information in a text
Can read a text closely
Can understand how the paragraphs in a text relate to each other
Can link simple information from different parts of a text
Can use the context to understand difficult parts of a text
Can draw simple conclusions

Can recognise and use basic grammatical structures/ function words in context

Mastery level 4
Pupils

Have a fairly wide vocabulary
Can work pout the meaning of unknown words from the context
Can understand quite complex sentences
Can understand quite long and complex texts
Can link information from different parts of a text
Can draw conclusions
Can make choices between some grammatical structures/ function words in order to express him/herself.

Mastery level 5
Pupils

Can use appropriate reading strategies
Have a quite wide and sophisticated vocabulary
Can understand complex sentences
Can understand long and complex texts
Can read between the lines and draw advanced conclusions
Can make choices between a range of grammatical structures/ function words in order to express him/herself.