Assessing texts written for a young, authentic audience, rated by peers and an authentic audience
Table of contents

Abstract

List of tables

List of figures

1. INTRODUCTION

1.1 RESUMÉ OF THE METHOD AND THE RESULTS OF MY FORMER THESIS

1.2 RESUMÉ OF THE LITERATURE REVIEW OF MY FORMER THESIS

2. LITERATURE REVIEW

2.1 WRITING TASKS IN EDUCATIONAL SETTINGS

2.2 INFLUENCES ON TEXT QUALITY

2.3 ASSESSING WRITING

3. Method

3.1 Research questions

3.2 Research groups

3.3 Research design

3.4 Variables

3.5 Materials

3.6 Operationalization

3.7 Analysis
3.7.1 Quantitative analyses ............................................................................................................. 46
3.7.2 Qualitative analyses .................................................................................................................. 49

4. Results ............................................................................................................................................. 50

4.1 Introduction ................................................................................................................................. 50
4.2 Perception of a suitable text for children .................................................................................. 50
4.3 Agreement on the Menno Steketee text .................................................................................... 51
4.4 Inter-rater agreement .................................................................................................................. 53
   4.4.1 Concept texts expert raters .................................................................................................. 54
   4.4.2 Concept texts grade 7 students ........................................................................................... 55
   4.4.3 Concept texts expert raters and grade 7 students ............................................................... 55
   4.4.4 Final texts expert raters ...................................................................................................... 56
   4.4.5 Final texts grade 7 students ............................................................................................... 57
   4.4.6 Final texts expert raters and grade 7 students ................................................................. 57
4.5 Contributions of the traits to the final grade ............................................................................. 59
   4.5.1 Correlations between the traits in the assessment models ................................................. 60
   4.5.2 Contributions of the traits to the final judgement ............................................................... 62
4.6 Improvement from concept to final text .................................................................................... 65
4.7 Contribution of revision and feedback on text quality .............................................................. 67
   4.7.1 Concept version .................................................................................................................. 67
   4.7.2 Revision ............................................................................................................................... 67
   4.7.3 Feedback and revisions ....................................................................................................... 68
   4.7.4 Implementing feedback ....................................................................................................... 68
4.8 Revision strategies freshmen students ...................................................................................... 69
4.9 Anecdotal material of writers revising ....................................................................................... 70

5. Discussion ....................................................................................................................................... 73

5.1 Introduction ................................................................................................................................. 73
5.2 Summary results .......................................................................................................................... 73
5.3 Discussion of content ................................................................................................................. 76
5.4 Methodological discussion ........................................................................................................ 77
5.5 Societal relevance ....................................................................................................................... 79
5.6 Future research ........................................................................................................................... 79

References .......................................................................................................................................... 81

Appendices ......................................................................................................................................... 86
Appendix 1: Writing assignment children’s page (Dutch) ................................................................. 86
Appendix 2: Feedback form grade 7 students (in Dutch) ................................................................. 87
Appendix 3: Instruction for writing a peer review (in Dutch) ............................................................ 89
Appendix 4: Three articles written by Menno Steketee, including Domestic viruses (in Dutch) .... 90
Appendix 5: Assessment model grade 7 raters and expert raters (in Dutch) .................................. 93
Appendix 6: Codebook SPSS ........................................................................................................... 96
Appendix 7: CD-ROM spss data ...................................................................................................... 101
No prior research on writing assessment focuses on assessing texts written for a young, authentic audience. Thus, an aim of this study was to develop an assessment model to assess these types of texts. The results showed that not under- or overestimating your audience is one of the most important features in an assessment method to assess texts written for a young, authentic audience. The assessment method then should question if the language and the content is appropriate for this particular audience.

This developed assessment method was used to actually assess texts by expert raters and the authentic audience (12/13 year old children). In prior literature nothing is known about the rating practices of young children, and this was therefore studied in this thesis. The results showed that the expert raters and the authentic audience differed in focal points when rating. The authentic audience seemed to focus more than the expert raters on the introduction, the lay-out, the spelling errors and punctuation in the text. The expert raters concentrated more on the higher order concerns. This study also suggested that the expert raters were more able to rate using the assessment model, while the authentic audience seemed to base her decision more on aspects outside the assessment model. The authentic audience also seemed to have more difficulties distinguishing between the various traits used in the model.

A study by Sato and Matsushima (2006) indicates that interaction with an authentic audience is beneficial for text quality. In their study the students received feedback orally, but feedback can also be provided in written form. There are no studies on the effect of receiving written feedback from the real readers on the quality of written texts. This thesis also aimed to investigate this gap in the literature. Freshmen students therefore received written feedback from their peers and from an authentic audience and then rewrote their drafts. The results suggest that implementing feedback from the authentic audience is beneficial for text quality, whereas implementing more peer feedback could not predict text quality.

The last sub study in this thesis was revisions, as there is no agreement between researchers whether revising positively or negative affects text quality. In this study revising more (or less) did not affect the final judgement of the text. This result was found as the students in this study hardly revised their texts, which decreases the possibility to find an effect of revision on text quality. If the students revised, they revised locally and used simple operations, as they are probably not able yet to revise globally and use more complex operations.
## LIST OF TABLES

**Table 1.** An example of a holistic scoring scale in ESL for the TOEFL Writing Test with six traits, only three traits are displayed here (Weigle, 2002, p. 113). ...22

**Table 2.** An example of an analytical scale borrowed from Tedick (2002, p. 33) ...25

**Table 3.** An example of a primary trait scale, borrowed from Tedick, 2002, p. 36. ...28

**Table 4.** Four assessment scales (Table based on Weigle, 2002, p. 109). ...28

**Table 5.** The activities in this study ...39

**Table 6.** Band scores and the total score on the text by Menno Steketee, by two expert raters and two grade 7 student raters ...52

**Table 7.** Inter-rater agreement of the ratings on the concept and final texts in the expert raters group and in the grade 7 students group and also between the expert raters and the grade 7 students, in percentages. ...59

**Table 8.** The prediction of each text aspects to the final judgement of the text, when the texts were rated by the expert raters ...62

**Table 9.** The prediction of each text aspect to the final judgement of the text, when the texts are rated by the grade 7 raters ...63

**Table 10.** Difference in judgement of the improvement of the texts (from concept to final text) by the raters (two expert raters and two grade 7 student raters per text). ...66
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Feedback model of Nelson and Schunn. The dotted lines imply negative relations, while the continuous lines indicate positive relations (Nelson and Schunn, 2009, p. 377).</td>
<td>13</td>
</tr>
<tr>
<td>Figure 2</td>
<td>A taxonomy of the types of revisions (Faigley and Witte, 1981, p. 403).</td>
<td>19</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Ratings on the concept texts by both types of raters.</td>
<td>56</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Ratings of the final texts by both types of raters.</td>
<td>58</td>
</tr>
</tbody>
</table>
1. INTRODUCTION

Audience awareness is an important quality for a competent writer (Berkenkotter, 1981). To develop audience awareness in educational settings, students are usually asked to make a list of features of their prospected readers. These lists contain information about, for instance, age, status and interests. Generally, a writer portrays one standard audience that has no individual differences: the writer then will not succeed in meeting the needs of all his readers. To enable writers to get a clear image of an audience with different readers, writers can be asked to interact with the actual readers: the authentic audience (Reiff, 2002). Students, however, are hardly ever given the chance to interact with an authentic audience (Rijlaarsdam et al., 2009), even though interaction with real readers improves the quality of texts (Sato and Matushima, 2006). Not only interaction, but also receiving feedback from an authentic audience will possibly be beneficial for text quality. I could, however, not find any studies about the influence of receiving feedback from an authentic audience on text quality. Also, writing for a young audience could be beneficial for text quality. In a study by Gunel, Hand and McDermott (2009) students were asked to explain difficult biology concepts to a young audience, with as a consequence that these students understood the biology concepts better than the students who wrote for a different type of audience (peers, teacher or parents). This study could suggest that writing for a young audience also has a positive effect on writing quality. I could, however, not find any studies that investigated this relation. This thesis aims to fill all these voids.

In my former Master Dutch Language and Culture, I wrote a thesis about receiving feedback on an article for a children’s page of a newspaper. This text was written by college freshmen. These students received feedback from their peers (18/19 year old students) and from an authentic audience (grade 7 students, 12/13 year of age). I studied the differences between these two groups in providing feedback, I examined the perception of the feedback by the students, and whether the freshmen implemented the feedback of the peers or the authentic audience more frequently. In the current thesis I will build on this former thesis and will study the data further. As the quality of the concept texts and the final drafts of the articles were not examined in the former study, this was investigated in the current thesis. The focus of the current study will be fourfold; firstly, it will attempt to develop an assessment model to assess popular science texts written for a young audience, secondly, it aspires to examine the relation between text quality and receiving feedback from both types of feedback providers, thirdly, the current thesis aims to compare both rating groups in the way they have assessed the texts and finally, the revisions made by the freshmen students were analyzed. This leads to the following research questions:

1) What is, according to the research literature, the most valid and reliable method to assess popular science articles for a young audience?

2) Which traits contribute most to the final judgement of the text, as rated by the expert raters and the grade 7 students?

3) What is the relationship between the final judgements of the texts and the (implementation of the) feedback received from two types of feedback providers?
Not only feedback has been studied, also revisions were studied, to gain some insight into how the students have revised their texts, but also to study the relationship between the number of revisions and text quality.

4) What are the features of the revisions made by the freshmen students and what is the relationship between the number of revisions and the final judgements of the texts?

However, before these questions can be answered, it is necessary to provide some background information about the former study. Thus, in the remainder of this introduction I will briefly recapitulate the results of my former thesis, and I will also summarize the literature of my prior thesis to provide a sound foundation for the current thesis.

1.1 Resumé of the Method and the Results of My Former Thesis

Freshmen in an Academic Writing course at the University of Groningen had to write an article for the children’s page of a newspaper. In this assignment (see appendix 1 for the assignment) they had to transform content from academic texts into an interesting text for children. The learning aim of this task for this Academic Writing course was to teach the students to take their audiences into account. For the purpose of this study, the students received feedback from 12/13 year old children in pre-university education. They also received feedback from their peers.

The children (grade 7 students) were asked to provide feedback using a form, which requested the children to answer questions about the text, for instance: “what is your opinion regarding the title”. For every question they had to indicate how well the student did on a Likert-scale ranging from 1, not good at all to 5, very good. Subsequently they had to substantiate their choice (see appendix 2 for the feedback form). The peers, on the other hand, were requested to discuss positive features of the text they had read, but they also had to indicate what the writer could do to improve his/her text (see appendix 3 for the feedback instruction the peers received). This difference between peers and grade 7 students in the way they had to provide feedback – free feedback or feedback form – was inevitable. As the peers needed to learn in the Academic Writing course how to provide feedback without using a feedback form; the grade 7 students were unable to provide feedback constructively without using a form. A pre-test demonstrated that when these children were not leaded in their feedback, they merely stated that they either liked the text, or that they disliked the text, without substantiating their opinions. Therefore, it was necessary to guide the grade 7 students by using a detailed feedback form. The feedback from both the grade 7 students and the college freshmen was analyzed using the model of Nelson and Schunn (2009), which was adapted slightly (for more information, see the former thesis).

The analysis of the feedback suggested that there were considerable differences between the feedback of the peers and the children. When the authentic audience provided feedback, they mentioned problems significantly more than the peers. The peers’ feedback, on the other hand, included solutions; or the problem with the matching solution. The strategy of the peers to provide solutions, may have been deliberate, as the feedback they had written was also read by the teacher of the Academic Writing course, and they were presumably trying to show their teacher that they are capable of improving the texts of others.
Another difference was that the peers localized feedback more frequently than the authentic audience. This could suggest that the peers recognized that the feedback is read by the writer, and that indicating where the problem is located in the text could be helpful. The authentic audience presumably lacked this awareness, and hardly ever localized the problem or solution.

The peers and grade 7 students also seemed to differ in the scope of their feedback; the peers focused significantly more on word level than the authentic audience; the children concentrated on the text as a whole (structure, topic). This discrepancy can be accounted for, as the students probably did not want to burden their peers with higher order problems, which takes considerable time and effort to alter. The peers plausibly also commented much on word level, because they wanted to indicate that certain words were not suitable for the authentic audience. By pointing out the importance of not under- or overestimating your audience, the peers probably wanted to demonstrate to the instructor of the Academic Writing course that they are capable of taking their audience into account. The authentic audience mostly commented on the topic, usually explaining why they disliked the topic. Or they provided feedback on the typical structure, which was a side effect of the assignment. In this assignment the students were asked to imitate the style of a writer, Menno Steketee, who uses a characteristic type of structure. The topics of his paragraphs are not logically coherent, but are only indirectly related.

Also the way both the peers and the authentic audience cloaked their message seemed to vary. The peers reduced the impact of their message by using *hedges* or by employing a *mitigation-compliment*: complimenting the writer first and then indicating a problem: “I think the sentences can be read easily, however within the framework of this assignment, it would be better to break up a few sentences into shorter ones.” The grade 7 students, however, mostly used direct language and did not use veiled terms to communicate their messages: “It has been written very childlike and there are a lot of paragraphs where they should not be.” The reason why the peers felt obliged to decrease the effect of their message using *hedges* and *mitigation-compliments*, might have been due to the fact that they knew their peers and they were in other classes with them. Freedman (1992) states that students do not feel comfortable critiquing their classmates out of solidarity. The grade 7 students, on the other hand, had nothing to lose, as they stayed anonymous and did not know the writers.

This thesis also examined if the students perceived the feedback from the peers or the authentic audience as more helpful. Their answers in the reflection reports indicated that they agreed more with the feedback of their peers, but this difference was not significant. The students significantly more frequently understood the feedback of their peers, than the feedback of the grade 7 students. Also, they implemented the feedback from their peers significantly more often than the feedback from the authentic audience. Presumably, they trusted the skills of their peers in providing feedback more, as peers took providing feedback more seriously and had more experience in writing feedback reports, as students stated in their reflection reports.

In this thesis it was also studied whether implementation was not only affected by the feedback provider (peers or children), but also by certain features of the feedback. A one-way between groups analysis of variance was conducted to examine the influence of the feedback features on the agreement with the feedback, the understanding and the implementation of the feedback. Localizing feedback seemed to have a positive effect on implementation and understanding, and also affected the agreement. The positive effect of localization on agreement, understanding and implementation seems comprehensible, as the students could understand the feedback better, because the feedback providers demonstrated where the problem could be found. Scope also had a positive effect on the understanding, agreement and implementation. When the feedback providers provided feedback on the sentence level (wording, style), this appeared to increase the agreement; if the feedback was directed at
the paragraph level (examples, explanations) the students mostly disagreed with the feedback. The understanding and implementation of the feedback diminished, if the feedback was aimed at the paragraph level or at the text as a whole (topic, structure). The students presumably agreed with feedback at the sentence level, as these comments usually reviewed the ‘flow’ of a sentence; writers tended to agree with these small suggestions for improvement. The student implemented or understood the feedback on paragraph level less frequently, probably as these comments were mostly directed at examples in the text that were not suitable for the readers, or at explanations that were not clear. These comments were more subjective and more time-consuming to improve, that might have been the reason why students did not implement feedback on paragraph level. Also the feedback at the text level as a whole was not implemented, as these comments plausibly take effort and considerable time to alter.

In short, the feedback from the peers was understood and implemented more by the students, probably because they trusted the feedback from their peers more than the feedback from the 12/13 year old children. The feedback features also might have affected implementation: peers localized their feedback, which might simplify implementing feedback, as the writers were able to find the problem in their texts. The peers also commented mostly on lower order problems, which was presumably less time-consuming to implement than the problems the children provided feedback on: the structure and the topic of the text.

This thesis has shed some light on which feedback features a feedback report has to have, in order for the feedback to be implemented by the writer. However, this thesis did not analyse if implementing feedback has a positive effect on text quality. The current thesis, among other things, aspires to examine this relation. However, before describing the actual thesis, a summary of the literature of my former thesis is described, which will be the basis for the literature review of the current thesis.

1.2 RESUMÉ OF THE LITERATURE REVIEW OF MY FORMER THESIS

1.2.1 APPROACHES IN WRITING EDUCATION

In the last 25 years there has been a shift in writing education. Until the 1970s the focus was on teaching the features of effective texts. This approach was defined as the product approach. In this approach students were taught to write well using text examples; specific attention was paid to text structures and spelling and grammar conventions (Galbraith & Rijlaarsdam, 1999; Hyland, 2003). This product approach has been criticized; it was claimed not to lead to an original expression of thought, but to merely stimulating the transformation of examples (Badger & White, 2000; Galbraith & Rijlaarsdam, 1999). Tuffs (1993) also noted that the product approach focused on the end result and disregarded the process to arrive at the final text. These critiques caused a decrease in the application of the product approach in educational settings.

With the decline of the product approach the process approach emerged. The product approach did not disappear entirely; teachers still used the product approach, but the process approach was the dominant approach in the 1980s (Matsuda, 2003). The process approach concentrated on the writing process: how writers can control the writing process by employing writing strategies (Galbraith &
Rijlaarsdam, 1999). Critique on this approach was directed at the vagueness of the classroom practice (Tuffs, 1993) and on the disregard of the social nature of writing (Hyland, 2003a).

The call for focusing on the social aspect of writing, marked the beginning of the genre approach in the beginning of the 1990s. This approach was, according to Hyland (2003a) and Galbraith and Rijlaarsdam (1999), not an entirely new approach, but another execution of the existing process approach. Writing in this approach was not an individual process, but the writer was asked to consider various social contexts to achieve sundry aims (Hyland, 2003a). The discourse community the writer belongs to is very important for this approach: students had to learn the aims and conventions of their own community (Galbraith and Rijlaarsdam, 1999) and students were taught different genres (Hyland, 2003a).

1.2.2 AUDIENCE AWARENESS IN THE PROCESS AND GENRE APPROACH

The difference between the process approach and the genre approach also became apparent in the perception of the audience. In the process approach, developing audience awareness was considered to be essential (Berkenkotter, 1981). Also a study of Rafoth (1985) suggested that audience awareness was an indispensable skill for competent writers: more competent writers had a detailed picture of their readers; less competent writers, however, merely depicted the main outlines of their audience. During the process approach teachers tried to enhance the audience awareness of their students; asking them to make a list of characteristics of the audience, which for instance described the opinions, social and economic status, interests and (background) knowledge of their prospected readers: the audience analysis (Reiff, 2002).

The practice of analysing the audience has been questioned by the end of the process approach (Reiff, 2002). According to Reiff (2002), the process approach lacked the interaction between the writer, the reader and the text. In the process approach writing was a process in which the reader hardly participated: the reader played a role at the start of the text during the audience analysis, and finally appeared again at the end of the writing process to read the text. The reader did not genuinely partake in the writing process, as the reader was put on the background during writing. Also the depiction of the audience was not necessarily correct (Reiff, 2002). In the genre approach, however, students were taught to write for multiple readers who all participated in one audience. Also the interplay of text, writer and reader was encouraged in this approach; asking students to interact with real readers, the authentic audience, through, for example, feedback. A pioneer study by Sato and Matsushima (2006) indicated that interacting with the audience had a positive effect on text quality. Not only interaction with the readers had positive effects, also taking a younger reader into account seemed to be beneficial. Gunel, Hand and McDermott (2009) found a positive influence on learning when students were asked to write for a young audience, as opposed to their peers, their teacher or their parents.
When studying feedback, the focus in studies lays mostly on perception, implementation, transfer or the quality of the texts that are written on the basis of the feedback (cf. Cho, Schunn & Charney, 2006; Nelson & Schunn, 2009; Bitchener, Young & Cameron, 2005). The nature of feedback can be analyzed using different kinds of typologies (cf. Artemeva & Logie, 2002; Beason, 1993; Nelson & Schunn, 2009; Tseng & Tsai, 2007). The following three categories are mentioned in most typologies: scope of the comment (local or global), focus of the feedback (wording, structure, content) and finally the function of the feedback (evaluation, suggestion or reaction of the reader).

Nelson and Schunn (2009) use two of these three categories in their model: scope and function (see: figure 1). Function in their model is used more specifically as affective language. Scope is placed in the cognitive part of the model. Other features in their model are: summarization, specificity and explanations. Nelson and Schunn then distinguish between four categories in the cognitive part of the model (1) summarization, (2) specificity, (3) explanations and (4) scope. These four categories in the cognitive part of the model steer, according to Nelson and Schunn (2009), the understanding of the feedback. The affective part of the model is made up of affective language. Affective language directs the agreement with the feedback. For a writer to implement feedback, he has to understand it and agree with it according to the model of Nelson and Schunn (2009).

Figure 1: Feedback model of Nelson and Schunn. The dotted lines imply negative relations, while the continuous lines indicate positive relations (Nelson and Schunn, 2009, p. 377).
Perceived helpfulness of the feedback mediates between feedback and implementation (revision), in other words, feedback that is perceived negatively by the writer, is less effective (Atwater, Waldman, Atwater & Cartier, 2000; Rucker & Thompson, 2003). Whether the feedback is perceived positively or negatively depends on the message that is communicated through the feedback, and on the image the writer has of the feedback provider (Ilgen, Fisher & Taylor, 1979). Ilgen, Fisher & Taylor even claim that the acceptance of the feedback is mainly subject to the messenger. They state that expertise is the most significant factor for writers to either accept or reject feedback. Therefore, feedback from teachers is accepted, as teachers are perceived by their students as competent and experienced, however, feedback from their peers is usually not taken seriously and is called into question (Cho, Schunn & Charney, 2006). Anecdotic studies show that students doubt that their peers take providing feedback seriously and they feel that their peers lack experience in writing feedback (Artemeva & Logie, 2002). Moreover, students are insecure about their own abilities to provide feedback to their peers (Van Gennip, Segers & Tillema, 2010). Even though students are reluctant to receive and provide feedback, Cho, Schunn & Charney found that there are no significant differences between feedback from peers and teachers.

1.2.5 IMPLEMENTATION OF FEEDBACK

For writers to implement feedback, is thus dependent on the perception of the feedback provider and the feedback message. Nelson and Schunn (2009) describe three more factors that could mediate between feedback and implementation, according to prior research: understanding, agreement and memory capacity. Nelson and Schunn claim that if feedback is understood, this enhances implementation. They found that feedback is better understood if a solution is given to the problem, if the problem is localized and if a summary is provided. They claim that if either of these three features are used, the feedback is more frequently implemented. They also found that providing explanations decreases the understanding of the feedback (Nelson and Schunn). Agreement is also beneficial to implementation according to Ilgen, Fisher and Taylor (1979), however, Nelson and Schunn have not found any indication in their study that agreement mediates between feedback and implementation. Memory capacity did play an important role in feedback, as implementing feedback demands considerable cognitive capacities. Nelson and Schunn, however, claim that memory capacity is more significant in oral feedback than in written feedback, and it was therefore not necessary to include it in their feedback model.

Even if the writer implements the feedback, it is not necessarily processed well. Sadler (1989) claims that there are three significant conditions that enable writers to benefit from feedback. Writers have to know: 1) what a good performance is (what is the aim or standard of this particular assignment), 2) what the relation of the current performance is with a good performance (ability to compare the current performance level with a good performance level) and 3) what they have to do to fill the void
between the current performance and a good performance. This ability is quite difficult for students, Yorke (2003) therefore claims that the skill in evaluating texts should be ameliorated in students.

In this summary of the literature I have described writing approaches, audience awareness and various aspects of feedback. This information will serve as background information and will be built on in the current thesis. The next chapter will review the literature with an emphasis on writing assessment models and text quality, but will start with some information about authentic writing tasks, writing for a young audience and sources that can influence text quality.

### 2. LITERATURE REVIEW

#### 2.1 WRITING TASKS IN EDUCATIONAL SETTINGS

##### 2.1.1 INTRODUCTION

Writing tasks in educational settings are usually unauthentic: the task is only meaningful in the school context. Researchers therefore encourage teachers to use authentic tasks in the classroom. This paragraph will describe the definition of authentic tasks, the aspects that need to be considered when developing an authentic task and the benefits of letting students engage in an authentic writing task.

In this thesis the students were asked to write about a scientific topic for an authentic audience. The audience was, however, not only authentic, the readers were also quite young, which required the writers to simplify scientific information. The second part of this paragraph will describe prior studies on writing for a young audience.

##### 2.1.2 AUTHENTIC WRITING TASKS

#### 2.1.2.1 DEFINITION OF AUTHENTIC WRITING TASKS

The definition of authenticity in learning is remote from uniformity. Shaffer and Resnick (1999) recognize this problem and have investigated the definitions of authenticity in learning in many studies and come to the conclusion that four dimensions are mentioned in authentic learning in for instance mathematics, history and language. 1) real-world authenticity; 2) authentic assessment; 3) personal authenticity and 4) disciplinary authenticity. The first dimension requires texts and assignments that could also emerge in real-life. Authentic assessment entails that during the assessment, students for instance should not only show what they have learned, but they should also learn something during the assessment. Personal authenticity involves engagement of the student with the task: learning a task should be significant for the learner. The fourth authenticity form is disciplinary authenticity, which asks students to act like a professional within their own discipline; building on prior research.
Duke, Purcell-Gates, Hall and Tower (2006) define *authenticity* not broadly in all learning contexts, but view it more narrowly; developing literacy in reading and writing. For a reading or writing task for them to be *authentic*, it has to comply with two characteristics: *purpose* and *text*. The purpose of an authentic task is that it has an actual communicative aim. Thus, the resulting product of a writing task should be read by the actual reader who requires that information or is interested in that topic. Reading with a purpose entails that students read information they actually need. The other dimension, *text*, means that the materials used in the educational setting can also be found outside the classroom or is specifically developed for that class.

Lidvall (2008) defines *authenticity* only for writing and mentions five aspects that are usually present in authentic writing: 1) the possibility for writers to choose their own topic; 2) real readers reading the text; 3) developing an actual product for the audience; 4) learning to write in various genres and 5) writers must see a purpose and a context in their writing task. A good example of an authentic writing task is, according to Lidvall, writing a letter to a real person and actually posting it.

### 2.1.2.2 BENEFITS OF AUTHENTIC WRITING TASKS

The tasks that students engage in in educational settings are very rarely found in the real world; these assignments usually have no purpose beyond the school context (Forte & Bruckman, 2006). Duke, Purcell-Gates, Hall and Tower (2006), however, claim that language is acquired best when fitted into a real situation. Also Delpit (1992) argues that discourses can merely be learned through real life encounters with that particular discourse. Using authentic tasks is not only necessary, but also has beneficial effects on the writers. Lidvall (2008), for instance, claims that the possibility for students to choose their own writing topics, and the visible aims in authentic writing tasks enhances motivation and interest. Duke, Purcell-Gates, Hall and Tower (2006, p. 354) notice that students “came alive” when they were writing texts for a real audience that will actually read their texts.

Purcell-Gates, Degener, Jacobson and Soler (2002) studied the effect of authentic reading and writing tasks in the adult literacy students classroom. They found that the adults who were engaged in highly authentic tasks, read and wrote more frequently out-of-school and also read and wrote more complex texts than students who had to immerse in literacy activities that were less authentic. If students participated longer in these classes the relation between authenticity and benefits was greater. In a follow-up study, Purcell-Gates, Duke and Martineau (2007) found similar benefits for authentic writing and reading tasks. Second and third grade students who received authentic informational and procedural texts in science, understood the texts better and were more able to write in this genre than the students who received low authenticity tasks.

Another form of an authentic writing task is publishing a text. Forte and Bruckman (2006) state that teachers and researchers recommend publication as a means to increase authenticity in writing tasks. To provide students with a real audience Forte and Bruckman developed a writing assignment for their undergraduate students that required them to publish on Wikipedia. This pilot study suggests that writing on Wikipedia made students regard writing as an interactive process that involves the writer and an audience.
As already described in the introduction of this thesis, the study by Gunel, Hand and McDermott (2009) suggests that writing texts for a young audience is beneficial for writing-to-learn: students who wrote an explanation on a biology topic for a young audience understood the science concepts better than students that had written for their teacher, peers or their parents.

Kroll (1985) has also studied the effect of writing for a young audience, but did not study the link with writing-to-learn, but focused on the way various age groups rewrite a complex text for a young audience. Students from grade 5, 7, 9, and 11, and freshmen, had to rewrite a difficult narrative for an audience of grade 3 students (8-9 year old children). Kroll immediately pointed out a problem that could occur during this assignment, namely that the participants would not sufficiently comprehend the original complex text to be able to rewrite it for a younger audience. He obviated this problem by not only letting the subjects read the text, but also by showing them a film of the narrative and offering them help when required. The results of this study reveal that the students from all grade levels decreased the complexity of the narrative; however, the older the student, the simpler the text. The older students also rewrote bigger parts of the text, focusing on the meaning of the text; while the younger students tended to concentrate on a more local level: on the substitution of certain words.

The study by Kroll indicates that students of different ages use different strategies for rewriting complex texts for a young audience. Gunel, Hand and McDermott’s study suggests that writing for a younger audience can be beneficial for learning complex concepts. There are, however, no other studies on the effects of writing for a younger audience on text quality. This thesis aims to shed some light on how writing for a young authentic audience affects text quality.

2.2 INFLUENCES ON TEXT QUALITY

2.2.1 INTRODUCTION

This thesis studies the effect of receiving feedback from an authentic audience and peers on text quality. This relation is, however, not straightforward. There are several mediating factors, apart from receiving feedback that can for instance influence text quality: writing beliefs, writing apprehension, writing self-efficacy, writing approaches and revision.

The first paragraph will shortly describe writing beliefs, writing apprehension, writing self-efficacy and writing approaches and their consequences on writing performance. The second paragraph will describe revision: the types of revisions that can be employed by writers and the impact of revision on writing quality will be elucidated.
2.2.2 INTRINSIC INFLUENCES

Personal beliefs of one’s competency to perform a certain task, which is called self-efficacy, a theory developed by Bandura (1977), is said to affect text quality (Zimmerman and Bandura, 1994). A study by Shell, Murphy and Bruning (1989) who studied the self-efficacy in reading and writing, also suggests that self-efficacy affects writing performance.

Writing apprehension also seems to affect writing performance (Daly, 1978). Writing apprehension involves the amount of anxiety students experience in their approach to writing. Some writers are not apprehended, while others try to avoid writing as much as possible, as they feel they would fall short in their attempt to do any writing (Daly & Miller, 1975a, 1975b; Faigley, Daly & Witte, 1981). Daly (1978) has studied writing apprehension as compared to writing competency for undergraduate students, using an indirect writing skill test, and he found that students with low writing apprehension scored significantly better on his test which involved for instance spelling, punctuation and case. The score of the students who had a low apprehension score, also scored significantly better on the test as a whole.

White and Bruning (2005) have also tested the beliefs of students and their relation to writing quality. They distinguish between transactional and transmissive beliefs in writing. If writers have high transmissive beliefs, they are minimally immersed in writing and do not see the writing process as an interactive process between writer, reader and text, but as a means to reiterate what has been said by other writers. Writers who mostly have transactional beliefs, however, are immersed in the writing process and their aim in writing is to personally engage in the writing process by putting their own opinions in the text. They found that students with high transmissive beliefs wrote texts with poorer structure and those essays were also of lower quality. The students who had high transactional beliefs wrote texts with better structure, voice, sentence fluency, content and conventions as punctuation, grammar, spelling and et cetera. And these students also wrote texts of higher quality.

The chosen writing approach can also affect writing performance. Lavelle and Guarino (2003) see writing strategy as the mediating factor between writing beliefs and writing performance. Thus, writing beliefs affect the strategy that is being used, and the chosen approach in turn influences writing quality. Writing strategies for instance involve planning, drafting and revising. Kellogg (1988) has found that the implication of certain planning strategies affects writing quality. He studied the influence of using pre-writing strategies and the effect on text quality. He found that especially (written and mental) outlining had a beneficial effect on the quality of the final text, as planning beforehand reduced the cognitive load of organizing during writing. Piolat and Roussey (1996) found similar results: they found that constructing an organized draft increased writing quality more than merely listing ideas.
2.2.3 THE INFLUENCE OF REVISION

2.2.3.1 TYPES OF REVISION

The writing task that the students had to perform in the current study requires revising after receiving feedback from their peers and the authentic audience. Revision is defined by Fitzgerald (1987) as changing parts in the text at any point during the writing process; these small or big alterations can change the meaning of the text or preserve it.

The most commonly used (basis) taxonomy for classifying different types of revisions is the model by Faigley and Witte (1981), see figure 2. They divide their model up into two types of changes: surface changes (left-hand side of the taxonomy) and text-base changes (right-hand side of the model). Surface changes involve changes that do not add or delete new information. Text-base changes do add additional information to the text, or delete paragraphs that change the meaning of the text. Surface changes are divided up into formal changes and meaning-preserving changes. Formal changes involve mostly edits in the text, while meaning-preserving changes entail for instance that language is substituted for similar content, but that the meaning of the text is preserved. The text-base changes are divided up into microstructure changes and macrostructure changes. This dichotomy distinguishes itself by the impact it has on the text: macrostructure changes alter the summary of the text, microstructure changes do not. But both types change the meaning of the text.

The revision types under meaning-preserving changes, microstructure changes and macrostructure changes involve additions, deletions, substitutions, permutations, distributions and consolidations. Most of these types are self-explanatory. However, the last three types need some elucidation. Permutations involve shifting parts of sentences or shifting with substitutions; distributions entail putting information that is placed in one sentence, into two sentences. In consolidations, information that was cased in two sentences, is now embedded in one sentence.

![Figure 2: A taxonomy of the types of revisions (Faigley and Witte, 1981, p. 403).](image-url)
Faigley and Witte (1981) have studied the changes made by six inexperienced writers, six intermediate students and six professionals. All six adult experts have experience in journalism. These types of writers seemed to differ in the ways they revise their texts. The experts employed text-base changes 34% of the time; the intermediate students used this category in 24% of all revisions. The inexperienced writers, however, changed the meaning of their texts only in 12% of all revisions. The inexperienced writers focused mostly on surface changes. These changes involve among other things, spelling, tense, punctuations, substitution of synonyms, deletions and additions in which the meaning of the text is preserved. A study by Lai (1986) also seemed to suggest that student writers mostly employ simpler operations: additions, deletions and substitutions. The expert writers in the study by Faigley and Witte (1981) mainly applied the following three types of revisions: addition, consolidation and distribution.

Not only the inexperienced writers revised on the surface, also the advanced students revised locally, namely 101 times per 1000 words. The experts also revised on the surface, but did this less frequently, namely 28.7 times per 1000 words (Faigley and Witte, 1981). Also Van den Bergh and Meuffels (2000) found that better writers focused on deeper aspects of the text when revising, and more competent writers made changes to bigger fragments of the text than less competent writers. Sommers (1980) found that less experienced writers concentrated on local aspects as grammar and spelling, while more experienced writers focused on local and global aspects of the text when revising. Also other studies found that students revise locally (Dave & Russell, 2010; Fitzgerald, 1987), focusing mainly on the word level (Lai, 1986). Sommers (1980) explains the tendency of students to revise locally as a lack of proper strategies: students have strategies to revise words and sentences, but they do not know how to revise the text as a whole. Wallace and Hayes (1991), on the other hand, claim that students focus on surface changes as they think that revision is merely changing the text locally. Faigley and Witte found that the intermediate and expert writers focused on surface changes after their last version of the text, to edit the last errors in their texts, while the inexperienced writers had already stopped revising the text at that point.

### 2.2.3.2 REVISION AND TEXT QUALITY

The relation between revising and text quality is not at all straightforward. Revision seems to depend on the initial quality of the text. Breetvelt, Van den Bergh and Rijlaarsdam (1994) state that revising a low quality text will deteriorate the quality of the text, while making changes in a good text will enhance the quality of the text. In their study they found that revising has a negative effect on text quality. A study by Van Gelderen (1997) also indicates that revising for a considerable number of writers does not increase the quality of the text, as compared to the initial text.

The number of revisions made by the writer seems also to affect writing quality. Even though one would expect revising more will have a positive effect on the final text, Stevenson, Schoonen and De Glopper (2006) found small evidence of the opposite: revising more had a negative influence on text quality. Their explanation of this finding was that writers who wrote a text of poor quality will have to revise more to make their texts better, while texts that were already of good quality, will merely need little adjustments. Van Gelderen (1997) found that better writers were also more able to revise than less competent writers.
2.3 ASSESSING WRITING

2.3.1 INTRODUCTION

In this study the concept texts and the final texts written by the freshmen students have to be assessed to investigate the differences between the peers and the authentic audience in assessing texts and to study if feedback and revision can predict text quality. To be able to execute this study, an assessment procedure has to be developed using former research on assessing writing.

Former research on writing assessment has been studied in various fields: it has a multidisciplinary approach. Huot (2002) describes the two main types of scholars working on writing assessment: composition scholars and educational measurement scholars. The composition scholars mostly concentrate on the context of writing, on reading and writing processes and on how teaching affects these processes. The educational measurement scholars have a more theoretical approach and look at how to sample texts, statistics and reliability and validity. Huot (2002, p. 58) in his book advocates collaboration between the fields as it would ameliorate “knowledge and understanding and improve the assessments we can create”. Approximately a decade after publication of Huot’s book there is no evidence of collaboration between those two fields (Behizadeh and Engelhard, 2011). In this literature review describing writing assessment, the research by both fields will be described: the theoretical aspects of measuring writing performance and the context of writing.

The literature review will mostly focus on different views of how to develop a proper writing assessment procedure. It will discuss various writing assessment methods: holistic ratings, analytic ratings, primary trait, multiple trait, prominent feature analysis, benchmark assessment and comparative methods. For every method the advantages and disadvantages for using each method will be outlined. Choosing a method is usually a means of either sacrificing reliability or validity. After deciding on a method, the researcher (or teacher) has to develop the writing assignment, the assessment scale and the scoring procedure. The literature review describes the aspects that should be carefully thought out when constructing this assessment. After developing the assessment, the texts are rated by raters. The last paragraph describes the variability between raters, especially novices and experts, how raters should be trained and the biases raters can have.

2.3.2 WRITING ASSESSMENT METHODS

Writing can be assessed in numerous ways. These methods are usually grouped into three classifications: direct testing, indirect testing and portfolio assessment. Direct testing refers to assessing writing products directly, thus rating the texts written by the writers, while with indirect testing, the writer does not write, but fills in a language test, usually by means of a multiple choice test. Portfolio assessment implies assessing multiple texts of a single writer (Weigle, 2002).

Direct testing is, according to Hamp-Lyons (2001), the oldest assessment method and can be traced back to China, at least to the Chou period; roughly 1000 years before Christ. Due to colonization, the Chinese assessment method was also adopted in Europe. In this method essays were assessed by
multiple raters. From the 1950s there was a call for more reliable methods to test writing ability, as the enrollments at universities increased and written examinations became more common. This increase resulted in the use of indirect testing; putting more emphasis on reliability than validity. From the end of the 1970s the assessment method shifted back towards direct testing, as this was a more useful way of assessing texts according to language teachers (Knoch, 2009). Validity, instead of reliability, dominates this method. Commencing from the early 1980s, portfolios were also used to assess writing ability (Hamp-Lyons). Yancey (1999) states that in portfolio assessment both reliability and validity are significant features.

In the literature usually three direct assessment methods are mentioned: holistic ratings, analytic ratings and primary trait ratings. Using a holistic rating, a text is read using a scale with usually 5 or 6 traits explaining the level of each trait, the text is then assessed with a single score. An analytical rating consists of judging sub aspects of the text, also using a scale, which produces multiple scores, these ratings are usually added up to retrieve a global score (Weigle, 2002). And finally, a primary trait scale focuses on only one aspect of a text, for instance if the writer has persuaded the audience (Schoonen & De Glopper, 1992). This scale is task-specific and yields a single score. A fourth assessment method is also often referred to: multiple trait scoring (Weigle). This method is similar to the primary trait, as it is task-specific and focuses on a certain topic/text. The difference with primary trait is that multiple trait considers more dimensions, and thus, has multiple scores. These dimensions in the multiple trait form are added up, similar to the analytical scale, to yield a global score (Tedick, 2002).

2.3.2.1 HOLISTIC RATINGS

Holistic scoring is also known as impressionistic, global or integrative scoring (Coombe & Evans, 2000). In this method the rater provides an overall and quick rating of the text. The procedure of holistic assessment usually has the following procedure: one or more raters read the text and hold it against a scale that states which criteria a text has to meet to obtain a certain score. These criteria can include structure, language use or content. These scores generally have 4 or 5 traits, with, for instance, a 1 for poor and a 5 for excellent quality. Each of these scores is predominantly exemplified with a benchmark essay. See table 1 for an example of a holistic rating scale. The rater(s) then provide(s) an impressionistic judgement of the text. When two raters are involved in the rating procedure and they disagree, a third rater interferes. Those two or three ratings are subsequently added up and averaged (Hamp-Lyons, 1995; Weigle, 2002). Holistic scales are not task-specific and can be used for other writing assignments (Tedick, 2002).

Table 1: An example of a holistic scoring scale in ESL for the TOEFL Writing Test with six traits, only three traits are displayed here (Weigle, 2002, p. 113).

| 6 | An essay at this level |

22
- Effectively addresses the writing task
- Is well organized and well developed
- Uses clearly appropriate details to support a thesis or illustrate ideas
- Displays consistent facility in use of language
- Demonstrates syntactic variety and appropriate word choice though it may have occasional errors

<table>
<thead>
<tr>
<th>5</th>
<th>An essay at this level</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>May address some parts of the task more effectively than others</td>
</tr>
<tr>
<td>-</td>
<td>Is generally well organized and developed</td>
</tr>
<tr>
<td>-</td>
<td>Uses details to support a thesis or illustrate an idea</td>
</tr>
<tr>
<td>-</td>
<td>Displays facility in the use of language</td>
</tr>
<tr>
<td>-</td>
<td>Demonstrates some syntactic variety and range of vocabulary, though it will probably have occasional errors</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1</th>
<th>An essay at this level</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>May be incoherent</td>
</tr>
<tr>
<td>-</td>
<td>May be undeveloped</td>
</tr>
<tr>
<td>-</td>
<td>May contain severe and persistent writing errors</td>
</tr>
</tbody>
</table>
Holistic ratings are considered reliable, as it usually provides a reasonable inter-rater reliability (White, 1984; Tedick, 2002). As a holistic rating is intrinsically impressionistic, they require little time and effort to apply, and thus this method is inexpensive. Despite listing the limitations of holistic ratings, White (1984) favors holistic assessment over analytic ratings in his article Holisticism. He especially appreciates the ability of holistic ratings to assess the text as a whole and not as sub parts, which is, according to him, a more natural way of reading and rating a text. For that reason he advises against analytic ratings, as it cuts the writing up into pieces. He ends his article with a recommendation for holistic ratings (p. 409):

Even the meanest bit of halting prose, even the most down-trodden of our fellow creatures, deserves to be taken as a living and vital unit of meaning, an artistic and human whole, not merely as a collection of scraps and parts.

Other researchers point out the limitations of holistic ratings more severely. They are particularly opposed to the single score, as it does not provide detailed information about the ability of the writer. They claim that a holistic rating is not meaningful for writers who want to know what they did well, and what they could improve: a holistic rating lacks diagnostic information (Schoonen and De Glopper, 1992; Knoch, 2011). Hamp-Lyons (1995, p. 760-761) expresses this problem aptly and evocative: “A holistic system is a closed system, offering no windows through which teachers can look in and no access points through which researchers can enter.” The single score is also not representative for written composition, as a text consists of various different aspects that usually differ in quality; the style can be good, while the text riddles with spelling errors. A holistic rating could then paint a rose-tinted picture of the ability of a writer that bears little relation to reality, as the rating does not demonstrate that the writer is incapable of writing without spelling errors (Weigle). Hamp-Lyons (1995) therefore states that texts cannot be captured in only one rating. Another problem lies in the different judgements of the raters. Raters will have various focal points and base their judgement on different characteristics of the text. When the text is assessed, leaving only a single score behind; it is uncertain what the reasons of the raters were to give a certain score (White, 1984; Weigle, 2002). Rating holistically can also lead to a halo-affect: some irrelevant characteristics of the texts are given too much emphasis in the scoring. Markham (1976) showed that when raters had to rate holistically, an insignificant feature as handwriting could predict the holistic score. Even though White (1984) is enthusiastic about rating holistically, he also points out the limitations. He mentions the narrow significance of a holistic score, as it is only meaningful seen in the context of the other texts the text was compared to.

2.3.2.2 ANALYTIC RATINGS

Analytic ratings, also referred to as discrete point marking and focused holistic marking, do not result in single scores, but yield multiple scores, as a rater focuses on various aspects of the text. Raters can for instance focus on content, structure, grammar, mechanics and vocabulary (Coombe and Evans, 2000). See table 2 for an example of an analytical scale. The standard analytic rating process usually has the following procedure: one or more raters read the text several times to assess all the sub parts of the written composition. These sub criteria are then predominantly added up to obtain a global score; it is also possible to put more weight on certain aspects, to emphasize important features (Weigle, 2002).
Coombes and Evans claim that texts assessed with analytical scales yield lower scores than texts rated holistically, as raters focus specifically on many different aspects of writing and therefore are able to find more problems.

One advantage of analytic ratings has already been alluded to above, namely that it has the ability to provide detailed information about the ability of the writer, for instance about their use of vocabulary and grammar, but also about the content and organization of the text. These sub scores provide the writer with feedback about the strengths and weaknesses of his text. This diagnostic information has been mentioned most as the chief merit of analytic scales (Weigle, 2002; Hamp-Lyons, 1995). An analytic scale is also more suitable for training raters than holistic scales, as this scale is more informative and is therefore easier to use. Another advantage of rating analytically is that it can yield higher reliability scores than holistic ratings (Weigle).

Similar to holistic scales, also analytic scales are criticized. Researchers are mainly commenting on the sub scales used in analytic scales. Coombe and Evans (2000, p. 4) for instance claim that “writing is a whole, not just parts summed up”. Another problem with sub scores, lies in adding them up to yield one global score. Weigle (2002) argues that when you average several sub scores, you will be deprived of considerable diagnostic information that had just been obtained. White (1984), as an advocate of holistic ratings, also claims that if a rater focuses on separate aspects of a text, he will be distracted from the actual purpose of the text as a whole. White also disputes if it is even possible to construct an analytic scale. He argues that this is practically unfeasible, as it is severely challenging to develop a scale with sub scales, as raters disagree about sub skills in writing. Also the labels for traits in analytic scales are usually unclear; raters will have different interpretations for for instance the label ‘vivid’ and therefore assess a text differently (Weigle). A more practical problem of analytic scales is that rating analytically is extremely time-consuming: it demands a rater to read a text several times to assess every trait (Weigle).

**Table 2:** An example of an analytical scale borrowed from Tedick (2002, p. 33)

**CONTENT – 30 TOTAL POINTS POSSIBLE**

<table>
<thead>
<tr>
<th>Score</th>
<th>Range</th>
<th>Criteria</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>30 – 27</td>
<td>Excellent to Very Good — • addresses all aspects of the prompt • provides good support for and development of all ideas with range of detail • substantive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 – 22</td>
<td>Good to Average — • prompt adequately addressed • ideas not fully developed or supported with detail, though main ideas are clear • less substance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 – 17</td>
<td>Fair — • prompt may not be fully addressed (writer may appear to skirt aspects of prompt) • ideas not supported well, main ideas lack detailed development • little substance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 – 13</td>
<td>Poor — • doesn’t adequately address prompt • little to no support or development of ideas • non-substantive</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ORGANIZATION – 20 TOTAL POINTS POSSIBLE**

<table>
<thead>
<tr>
<th>Score</th>
<th>Range</th>
<th>Criteria</th>
<th>Comments</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Criteria</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 – 22</td>
<td><strong>Excellent to Very Good</strong> — • great variety of grammatical forms (e.g., range of indicative verb forms; use of subjunctive) • complex sentence structure (e.g., compound sentences, embedded clauses) • evidence of “Spanish-like” construction • mastery of agreement (subj/verb; number/gender) • very few errors (if any) overall with none that obscure meaning</td>
<td></td>
</tr>
<tr>
<td>21 – 18</td>
<td><strong>Good to Average</strong> — • some variety of grammatical forms (e.g., attempts, though not always accurate, of range of verb forms, use of subjunctive) • attempts, though not always accurate, at complex sentence structure (e.g., compound sentences, embedded clauses) • little evidence of &quot;Spanish-like&quot; construction, though without clear translations from English • occasional errors with agreement • some errors (minor) that don’t obscure meaning</td>
<td></td>
</tr>
<tr>
<td>17 – 11</td>
<td><strong>Fair</strong> — • less variety of grammatical forms (e.g., little range of verb forms; inaccurate, if any, attempts at subjunctive) • simplistic sentence structure • evidence of &quot;English-like&quot; construction (e.g., some direct translation of phrases) • consistent errors (e.g., with agreement), but few of which may obscure meaning</td>
<td></td>
</tr>
<tr>
<td>10 – 5</td>
<td><strong>Poor</strong> — • very little variety of grammatical forms • simplistic sentence structure that contains consistent errors, especially with basic aspects such as agreement • evidence of translation from English • frequent and consistent errors that may obscure meaning</td>
<td></td>
</tr>
</tbody>
</table>

**LANGUAGE USE/GRAMMAR/MORPHOLOGY – 25 TOTAL POINTS POSSIBLE**

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Criteria</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 18</td>
<td><strong>Excellent to Very Good</strong> — • well-framed and organized (with clear introduction, conclusion) • coherent • succinct • cohesive (excellent use of connective words)</td>
<td></td>
</tr>
<tr>
<td>17 – 14</td>
<td><strong>Good to Average</strong> — • adequate, but loose organization with introduction and conclusion (though they may be limited or one of the two may be missing) • somewhat coherent • more wordy rather than succinct • somewhat cohesive (good use of connective words)</td>
<td></td>
</tr>
<tr>
<td>13 – 10</td>
<td><strong>Fair</strong> — • lacks good organization (no evidence of introduction, conclusion) • ideas may be disconnected, confused • lacks coherence • wordy and repetitive • lacks consistent use of cohesive elements</td>
<td></td>
</tr>
<tr>
<td>9 – 7</td>
<td><strong>Poor</strong> — • confusing, disconnected organization • lacks coherence, so much so that writing is difficult to follow • lacks cohesion</td>
<td></td>
</tr>
</tbody>
</table>

**VOCABULARY/WORD USAGE – 20 TOTAL POINTS POSSIBLE**

<table>
<thead>
<tr>
<th>Score Range</th>
<th>Criteria</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 18</td>
<td><strong>Excellent to Very Good</strong> — • sophisticated, academic range • extensive variety of words • effective and appropriate word/idiom choice and usage • appropriate register</td>
<td></td>
</tr>
<tr>
<td>17 – 14</td>
<td><strong>Good to Average</strong> — • good, but not extensive (less academic), range or variety • occasional errors of word/idiom choice or usage (some evidence of invention of &quot;false&quot; cognates), but very few or none that obscure meaning • appropriate register</td>
<td></td>
</tr>
</tbody>
</table>
2.3.2.3 PRIMARY TRAIT

A primary trait scoring concentrates on one global feature of the text (Tedick, 2002). See table 3 for an example of a primary trait scale. This method has been developed to assess how well students write in certain genres; for instance persuasion (Weigle, 2002). Primary trait scoring can be reasonably reliable if the assessors are properly instructed (Cooper & Odell, 1977). The advantage of primary trait scoring is that teachers solely concentrate on one aspect of writing ability (Tedick). This advantage is according to Schoonen and De Glopper (1992), a disadvantage, as using a primary trait scale does not provide much diagnostic information about writing ability: a text that has obtained a high score can still contain many spelling errors. Constructing a primary trait scale also requires considerable time and effort, with as a consequence that this method has not been applied frequently in educational settings (Weigle).
Table 3: An example of a primary trait scale, borrowed from Tedick, 2002, p. 36.

<table>
<thead>
<tr>
<th>Primary Trait: Persuading an audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 — Fails to persuade the audience.</td>
</tr>
<tr>
<td>1 — Attempts to persuade but does not provide sufficient support.</td>
</tr>
<tr>
<td>2 — Presents a somewhat persuasive argument but without consistent development and support.</td>
</tr>
<tr>
<td>3 — Develops a persuasive argument that is well developed and supported.</td>
</tr>
</tbody>
</table>

2.3.2.4 MULTIPLE TRAIT

Multiple trait scoring is almost similar to primary trait, as it also focuses on one category of writing, for instance persuasion. However, in multiple trait scoring the scale focuses on more dimensions than one. It usually concentrates on three or four traits, for instance three traits focusing on audience (Tedick, 2002). The multiple trait scale seems to be similar to an analytical scale, the difference is that an analytical scale uses general traits as grammar, vocabulary and mechanics; the dimensions in a multiple trait scale, however, are task-specific. An overview of the differences between holistic scales, analytic scales, primary trait and multiple trait can be found in table 4.

Table 4: Four assessment scales (Table based on Weigle, 2002, p. 109).

<table>
<thead>
<tr>
<th></th>
<th>Specific to a particular writing task: narrowly defined</th>
<th>Generalizable to a class of writing tasks: broadly defined</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single score</td>
<td>Primary trait</td>
<td>Holistic</td>
</tr>
<tr>
<td>Multiple scores</td>
<td>Multiple trait</td>
<td>Analytic</td>
</tr>
</tbody>
</table>

2.3.2.5 PROMINENT FEATURE ANALYSIS

Prominent feature analysis has been developed by Swain, Graves and Morse (2010). In a prominent feature analysis, the raters construct the criteria for every writing task. The raters discuss the features, positive and negative, that attract attention: prominent features. An advantage of this method is that it is narrowly fitted to the writing assignment and focuses on “language and rhetoric” (Swain, Graves and Morse, p. 89). A disadvantage of this method is that it is takes a lot of time and effort to develop.
Benchmark or anchor assessment implies that raters compare texts with exemplars that were already assessed at an earlier stage. This scale can for instance represent the differences in quality. One of the first anchor scales has been constructed by Hillegas (1912, in Behizadeh & Engelhard, 2011). Anchor assessment is hardly ever used separately, but is usually applied in combination with other assessment methods. Feenstra (2012) for instance used benchmarks with analytical scales to assess content, correctness and structure. She found that using an analytical scale in combination with benchmark essays only enhances the inter-rater reliability significantly for the structure of the text and not for content or correctness.

**2.3.2.7 COMPARATIVE METHODS**

The method of *comparative judgement* was founded by Thurstone in 1927 and implies that texts are assessed through a pairwise comparison. Every rater receives pairs of texts and for every pair they have to indicate which text is of the highest quality (Politt, 2012). Advantages of this method according to Bramley (2007) are that assessing texts is easy and it is simple to employ. *Comparative judgement* is also a very reliable assessment method (Politt, 2012). A disadvantage *comparative judgement* shares with primary trait analysis, prominent feature analysis and analytic ratings is that it is a highly time-consuming method. Another disadvantage is that raters compare the same texts several times, and therefore could be biased in their new judgement, as they might remember what their decision was for the other text (Bramley, 2007).

Politt (2012) has altered the comparative judgement method into *adaptive comparative judgement*, which differs to *comparative judgement* in that the raters and the texts are not matched randomly, but the pairs are generated by a computer program. First, all texts are rated randomly using pairwise comparison. Second, when every text is assessed once, the texts that were assessed as best are offered by the computer in pairs and also the texts that have lost the competition are presented in pairs. This matching method continues for several rounds, matching up texts that are very similar in quality, acquiring considerable accuracy. This method obtains very high reliability statistics.

**2.3.3 DEVELOPMENT OF AN ASSESSMENT METHOD**

To develop an assessment that is valid (legitimate and meaningful), the constructor has to have a clear picture of the aspects that can affect validity. Weigle (2002) describes the following aspects: *the writing assignment, the text, the assessment method, raters’ features, writer’s features apart from their writing skill* and finally, *environmental factors*. These environmental factors involve the environment when the text was written (school or elsewhere). To control for these factors, the development of the writing task, the chosen assessment method and the scoring procedure have to be carefully thought out. In the
remaining paragraphs the construction of the writing task, the assessment scale and the scoring procedure will be described.

A writing assignment has to contain proper instructions. Bachman and Palmer (1996) focus on texts that students have to write under time pressure. And for this aim they describe which aspects should be included in the instructions. They recommend that the instructions should be understandable and not too long, as reading instructions should not take up too much of their time during the test, and finally the instructions have to be sufficiently specific. Weigle (2002) has added three more directions for the instructions of a writing assignment. She claims that the instructions have to include a description of the audience and the goals, it has to provide an approximate length of the text and students should know what the scoring criteria will be.

The topic of the writing assignment has to be attuned to the students and should not be too general, as this could elicit texts that are very alike. Personal topics are advised against, as some students feel uncomfortable writing about themselves, similarly, it is difficult to rate personal texts (Weigle, 2002). To ensure that students all have the same opportunities in writing a good text, they should be given background information about the topic (Weigle).

After the writing assignment has been made, the chosen assessment method has to be developed. Weigle (2002) describes the construction of an analytical scale. Weigle (2002) claims that an analytical scale should measure what it is supposed to measure, to ensure the validity of the assessment. Thus, an analytical scale should not be vague or implicit, and every rater should be able to understand and apply the analytical scale. When designing a scale, the following four aspects should be regarded: the user of the scale, so for instance, is the aim of the scale to provide diagnostic information to the writers or is the scale meant as an aid to simplify the rating task? The second aspect is the consideration of the significant features of writing that will be tested in the analytical scale, thus, will the emphasis of this writing task be on audience? And how many aspects on audience will be represented in the scale? The third aspect to regard is the number of scoring levels that can be applied in the analytical scale. The number of scoring levels that can be used in an analytical scale depends on the routine of the raters. The more skilled raters are able to distinguish between more levels than the less experienced raters. The last aspect that should be considered according to Weigle is the scoring procedure. The developer should consider whether the analytical scores on all aspects can be added up to obtain one score, or if the ratings on all the aspects are maintained and analyzed individually. Obtaining one score has the advantage that it is more reliable than several ratings. And using only a single score simplifies analyzing the quality of the texts more than studying all scores separately. A disadvantage of adding up the analytical scores to obtain a single score is how to calculate the final score. The developer of the assessment has to decide whether some scores should be given more weight. How to come to a weighted single score is extremely difficult and therefore Weigle (2002) advises to seek help from a statistician.

White (1984) provides more recommendations to get a high reliability score for an assessment method. He states that firstly the analytical scale has to describe the scoring procedure. And secondly, exemplars should be used for practicing the rating, so that there is agreement amongst the raters what a certain score on a text should look like. Thirdly, White (1984) advises that every text should be rated by two people, and if these raters disagree, a third rater has the deciding vote. Fourthly, every text should preferably be rated in the same room, to control for environmental variables. During the scoring procedure, a leader should be appointed who guides the rating process. White emphasizes that this leader should be competent, as his capacities as a group leader can highly affect the rating process, making a scoring procedure successful or unsuccessful. If raters are not only judges in this rating
process, but will be rating in the future, a report should be made with information about the consistency of these raters.

2.3.4 Raters Assessing Writing

2.3.4.1 Reliability

Reliability is not only an important feature to consider when choosing an assessment method, but also during the rating process. There are two significant aspects for the reliability of raters: intra-reliability (self-consistency) and inter-rater reliability (agreement between raters). Intra-reliability refers to the reliability of one rater: the score should be similar if a rater, for instance, rates two texts of the same quality. Inter-rater reliability implies the consistency in scoring of two raters on the same text (Weigle, 2002). The reliability can be evaluated using statistics and will lead to a correlation coefficient between 0 and 1, showing how strong the relation is between the scores. A score near 0 indicates that there is hardly any relationship between the scores, while a score near 1 indicates a strong relationship between the scores. If inter-rater reliability is low, this can be enhanced, firstly, by increasing the number of raters rating every text, secondly, by specifying the rating scales and thirdly, by training the raters (Weigle).

2.3.4.2 Rater Variability

One of the biggest problems with writing assessment is that writing is extremely complex, and the global impressions readers have of a text will implicitly influence their opinions. Therefore is it virtually impossible to completely discover what raters base their ratings on. Also Lumley (2002) claims that it probably is not feasible to try to study how raters come to their scores, as raters are presumably not aware of what they base their judgements on; they are most plausibly affected by a plethora of text aspects. Lumley did try to study rater variability in rating processes and he found that if raters are using an analytical scale, they will all differ in their focus on different scale descriptors.

Cumming (1990) tried to study the decision-making process in rating. And he found that when raters were asked to assess second language texts holistically, they all differed in their rating process. The difference was especially visible between novice and expert raters: experts reflected more on their assessment and counted the ideas represented in the text more frequently. Novices on the other hand were mostly editing and basing their judgements on intuition. When rater variability is studied, researchers usually focus on this strand of research: the differences between novice raters and expert raters. Weigle (1999) for instance found that expert raters appreciated imaginative texts more than novice raters, who thought that some fragments were irrelevant, while experts, on the other hand, valued the original thought of the writer.

Schoonen, Vergeer and Eiting (1997) also studied the difference between experts and novices in rating texts elicited by different types of writing tasks. In one writing task, called the ‘specified task’, writers generated and structured content. In the other task, the ‘interlinear revision task’ writers
received a text that fulfilled the rhetorical goals, but that had to be revised. Schoonen, Vergeer and Eiting found that novices are less reliable than experts in rating the ‘usage’ (grammar, style and vocabulary) of a text, a term by Schoonen, Vergeer and Eiting, for the more free ‘specified task’. If the writers had to rate the ‘interlinear revision task’, both the novice and expert raters could rate ‘usage’ reliably. Both types of raters were able to rate the content of the text reliably for both types of writing assignments. A study by Shohamy, Gordon and Kraemer (1992) also shows that in some ways novices are able to assess texts reliably. They found that inexperienced raters were able to rate texts, provided that they are aided by an analytical scale.

Other differences between experienced raters and inexperienced raters is that experts rate more severely than novice raters on both analytical and holistic scales (Sweedler-Brown, 1985). Weigle (1998) however found that novices rate stricter than experts and her study showed that novices are less consistent in their ratings. Training both novices and experts seemed to affect the ratings, as after rating both types of raters rated more consistently, thus improving intra-rater reliability. Rater training, however, did not seem to enhance inter-rater reliability: there were still differences between raters in their tendency to rate more leniently or strictly.

2.3.4.3 RATER TRAINING

Weigle’s study (1998) could suggest the importance of rater training to improve the reliability of an assessment. Training raters usually involves raters being coached in understanding the criteria in the scale, thereby decreasing rater variability. In a training the raters are also explained what level the students are expected to have and the importance of agreement between the raters is stressed (Weigle, 1994). Another part of the training is that raters are familiarized with the scale by training them with exemple essays. Usually examples are chosen that are representative for various ability levels, but also unusual examples are used (Elder, Barkhuizen, Knoch and Von Randow, 2007). White (1984), as also described above, considers using examples an essential part of the training as it is an aid to reach a high reliability score.

However, some researchers argue that training raters is detrimental to the assessment process, as training could deteriorate the rating process, as raters rely on scoring criteria and are therefore not able to rate texts properly anymore (Huot, 1990). Charney (1984) claims that training in rating (holistically) is not beneficial for the rating process, as it makes raters focus on small, conspicuous, easy to rate, aspects of writing as spelling and grammar or handwriting, instead of the text as a whole, which is preferable. Weigle (1998) however argues that simplifying the scoring procedure is necessary for inter-rater reliability. When a text is rated as a whole, validity is taken into account, while researchers, who prefer to simplify the scoring procedures to obtain agreement among raters, choose reliability over validity. Obtaining total inter-rater reliability, thus similar severity among raters, is not feasible according to Weigle (1999) and Lumley and McNamara (1995), as lenience or strictness in rating is a personality trait that cannot be diminished by rater training. It is therefore advised to not use only one rater to assess a text, as one score mostly provides information about the severity of the rater and not the ability of the writer. McNamara and Lynch (1997) argue that using two raters enhances the reliability considerably and that three raters should be used for ratings the other two raters do not agree on.
Lumley & McNamara (1997) found another problem with rater training. Their study showed that raters who were trained on two occasions (18 months apart) would rate similarly on these times, but when they had to do the ‘actual’ rating, two months later, they rated differently. Lumley & McNamara then suggest that raters should refurbish the training before the actual rating process begins.

2.3.4.4 RATER-BIASES

When raters rate a text, the aim is not only to reach inter-rater and intra-rater reliability, but also that their scores are materialized objectively. This is, however, not possible. Ratings are subjective in nature and will be affected by external influences. A study by Grobe (1981) for instance reports that style, spelling, grammar, syntax, vocabulary and text length contribute to holistic scores. The best indicators for holistic scores are length and spelling. Thus, raters are focusing on sub features instead of the text as a whole, even though texts should be regarded as a whole and preferably not be affected merely by length or spelling. Focusing on one variable that affects the rating as a whole is a judgement bias called the halo-effect. This can for instance involve a judgement about the personality of the writer or about one specific feature of the text that affects the score on the text as a whole (Meuffels, 1994). Another example of the halo-effect is that some raters tend to give high-ability writers higher grades; as they are more willing to see problems in the text as unfortunate mistakes. However, other raters will score texts from good students lower, as they expect more of these writers (Schaefer, 2008).

Another effect that can affect ratings is the signification effect. This effect involves the differences between raters in what they find important in texts. One rater focuses mostly on content, while other raters for instance base their decision mostly on wording (Meuffels, 2002).

When more than one rater assesses the same text, all texts should be rated independently. One rater should not be affected by the score given by another rater. However, if a rater is affected by other raters’ comments or scores this is called the contamination effect bias (broad view) (Meuffels, 2002). Contamination effect bias (narrow view) involves that a rater has a personal advantage by rating a text in a certain manner. If a teacher is for instance judged by how many students receive positive results in his class, will this affect his manner of rating the texts of his students (Meuffels, 1994).

Raters are also biased by the sequence in which texts are rated: the sequence effect. If a rater has judged several high quality texts, the rater will have a tendency to judge a text of moderate quality as having low quality. Similarly, after rating a few bad texts a rater will probably judge a moderate text more positively (Meuffels, 1994). Another problem with rating several texts is that norm shift can occur. If the rater has judged several texts and notices that the quality of the texts in general is high; the rater unconsciously increases the level the texts should reach and therefore rates all the other texts more severely (Meuffels, 1994).

Personal comparison is the tendency of a rater to give certain scores. Some raters will be very strict, while others are lenient in their judgement. Another example of a personal comparison is the range of the grades: some raters will only give grades with very few differences; others will use the whole scale (Meuffels, 2002). Raters will also differ in the knowledge and expertise they have in a certain field. Raters who lack certain knowledge to properly rate a text will rate a text with mistakes in
that area higher than a rater who is able to spot these errors. This is called the *competency effect* (Meuffels, 1994).

In sum, to assess a text, a suitable assessment method has to be chosen, which depends firstly on the type of writing task, secondly, if diagnostic information should be given to the writers, thirdly on the amount of time available and finally, on the numbers of raters that can assess the texts. When designing an assessment method; a writing assignment, a scale and a scoring procedure should be constructed. In the scoring procedure, the raters are a significant part. White (1984) advises that at least two raters assess a text for a proper reliability score. Choosing certain raters should be a conscious one, as there is great rater variability, especially between novice and expert raters. Rater training is said to be an indispensable part of the scoring process, while other researchers claim that it is detrimental to the rating process. There seems to be agreement by other studies about the effect of rater training, indicating that rater training only enhances intra-rater reliability, and not inter-rater reliability. When setting up a scoring procedure also various rater-biases should be considered.

The next chapter will describe the method of this thesis and will also go back to the literature described here, as I will describe the scoring procedure I have used in this thesis. More specifically I will enunciate my considerations for choosing an analytical scale in combination with a multiple trait scale to assess writing quality; the specifications of choosing certain levels and traits in this scale; the writing assignment; the number of raters I have asked to rate a text and the details of the training of the raters.
3. METHOD

3.1 RESEARCH QUESTIONS

The literature review described the benefits of letting students write for an authentic audience, the many aspects affecting writing performance, including revision, and, finally, the complexity of assessing texts. Prior studies indicated that the perception of feedback affects the implementation of feedback and that interacting with an audience can improve text quality. No studies, however, investigated if receiving feedback from a young, authentic audience ameliorates texts. This thesis aims to investigate this gap in the literature. As was already explained in the literature review, raters vary in their focal points when rating a text. In former research the focus is mostly on ‘novice’ and expert raters. These novice raters are mostly students or new teachers rating texts. I could not find any information about the rating processes of young children. This thesis also aims to fill this void. The literature review also showed that the contribution of revision to text quality is not known; some researchers claim that revising more will deteriorate the text, while other researchers found that revising more decreased text quality. In this thesis the influence of revising more on text quality will be examined. Another goal of this study is to develop a reliable and valid method to assess texts. This leads to the following four research questions:

1) What is, according to the research literature, the most valid and reliable method to assess popular science articles for a young audience?

2) Which traits contribute most to the final judgement of the text, as rated by the expert raters and the grade 7 students?

3) What is the relationship between the final judgements of the texts and the (implementation of the) feedback received from two types of feedback providers?

Not only feedback has been studied, also revisions are studied, to gain some insight into how the students have revised their texts, but also to study the relationship between the number of revisions and text quality.

4) What are the features of the revisions made by the freshmen students and what is the relationship between the number of revisions and the final judgements of the texts?

This chapter will describe the subjects who participated in this study, the research design, the variables, the materials, such as the assignment and the development of the assessment model, and it describes the operationalization of the variables and finally it elucidates which statistics and/or methods are used to analyze the data.
3.2 RESEARCH GROUPS

3.2.1 WRITERS: FRESHMEN

The sample of this study consisted of freshmen and G7 students, both groups are statistically selected using accidental sampling. The freshmen were students Communication and Information Sciences at the University of Groningen. The sample consisted of two seminars of the course Academic Writing. 38 students have written a first draft of an article for a children’s page of a large national newspaper. After they have received feedback from their peers and the authentic audience they have written the final version of their text. The freshmen also had to specify in a reflection report what their perception of the received feedback was and they had to indicate which feedback they had implemented in their final texts. The specifics of this assignment will be laid out in section 3.5.1. Eventually the texts of 20 students (19 women (95%), 1 male (5%)) could be studied, as these students provided the final version of the article and all reflection reports, and received 2 peer feedback reports.

3.2.2 FEEDBACK PROVIDERS: AUTHENTIC AUDIENCE AND PEERS

Feedback was provided on 38 articles by the following sample: 50 G7 students (varying in age from 11 to 13 years) from two classes in pre-university education. This sample was selected using accidental sampling. These children have been chosen to provide feedback on the texts, because a prior study showed that providing feedback was too difficult for G6 students. These students for instance stated that certain words were too difficult for them. In addition, these students were not very proficient in providing feedback. Their comments were often limited to: “I liked it.” G7 students, who, in addition, were in this class for six months, were most likely to be better at providing feedback. The choice for pre-university education was based on the level of the texts written by the students. The students had to imitate the style of Menno Steketee, published on the children’s page of the Dutch newspaper NRC Handelsblad. The audience of this newspaper consists particularly of highly educated people (NRC Handelsblad, 2006). Thus, highly educated parents will probably read NRC Handelsblad. In this study children in pre-university education were examined, as this education level has the highest percentage of highly educated parents (Lange and Dronkers, 2007).

Feedback was not only provided by grade 7 students, also the peers of the freshmen students have given feedback on the concept texts. Every student received two feedback reports from two peers. Thus, the students received four feedback reports from two peers and two grade 7 students on their concept texts.
3.2.3 Raters: Experts and Authentic Audience

The concept texts and final texts of the freshmen were assessed by 7 expert raters, all female (varying in age from 21 to 26 years). These raters were all advanced or graduated students of the Faculty of Arts at the University of Groningen. Students from this Faculty were chosen, as these students are interested in language and texts, and usually enjoy reading. Therefore they usually are skilled readers and competent writers. This sample was also selected using accidental sampling, as these students responded to my call to assess texts. The 7 expert raters have rated 11 or 12 texts using an assessment form, which will be discussed in section 3.5.2.

The authentic audience also assessed the texts written by the freshmen. The authentic audience consisted again of students in the first class of pre-university education from exactly the same school as the authentic audience who has given feedback on the texts. As the assessment has been done a year later, no similar students have given feedback and have rated the texts; thus, the samples were independent. 46 G7 students (varying in age from 11 to 13) all have assessed 2 articles, 1 concept version and 1 final version of two different texts. This, however, meant that some articles were assessed more than twice. Because the texts were assessed by the peers also only twice, the surplus data was randomly deleted from the dataset.

3.3 Research Design

The research design of this study was correlational: this study does not contain experiments, it contains observations by the researcher. It examined the contributions of feedback and revision on text quality. Also the contribution of the traits in the assessment model to the final judgement was analyzed; in other words, how for instance structure contributed to the final judgement of the text. This thesis also studied the validity and reliability of the rating process. This process was examined by, firstly, investigating the perception of the raters of what an assessment model for a children’s article should look like (qualitative study); secondly, by analyzing the agreement between the raters in assessing the texts and finally, by examining the agreement on a text written by Menno Steketee. By evaluating the agreement on the ratings on the text by Menno Steketee the validity of the assessment model could be studied, as it was known that this text should yield a high score on the assessment model. Another part of this study was the investigation of the revisions by the students; this study was partly quantitative and partly qualitative. The quantitative part discussed some descriptive statistics of the revisions. The qualitative part described the revisions of two students when receiving feedback and how a writer could revise successfully or unsuccessfully. This last study provided a more meaningful, in depth insight in the writing products of students, when a writer is asked to implement feedback from two different sources.

The studies described above were chiefly quantitative in nature, however, two studies were not quantitative, as was already alluded to above: the perception of the raters of a suitable text for children and the description of the revisions of two students. More information about the analyses of all these separate studies can be obtained in 3.7.1 Quantitative analyses and 3.7.2 Qualitative analyses.
3.4 VARIABLES

The following variables were investigated in this study:

- Perception of what a suitable text is for 12/13 year old children
- Number of feedback comments
- Number of feedback implementations
- Number of revisions
- Revisions
- Text quality

3.5 MATERIALS

In this study various materials were used: the assignment used to write the text, the concept texts and final texts that have been scored, and the model to assess the texts. These materials will be discussed in this section. The last part of this section also describes the instruction for using the assessment model and the process used to assess the texts.

3.5.1 ASSIGNMENT

In this assignment, the freshmen had to transform a complex science text into a popular-science text for children of about 12 years old that could be published on the children’s page of the Dutch newspaper NRC Handelsblad. They had to imitate the style of Menno Steketee, who uses short sentences and paragraphs, humor and a typical structure. The freshmen were given some examples of the texts written by Menno Steketee to be able to imitate him more closely (see appendix 4 for texts written by Menno Steketee). It was emphasized strongly that they had to take their audience into account. In appendix 1 the instruction for the writing assignment can be found.

3.5.2 THE ASSESSMENT PROCESS

Assessing the quality of texts meant that an assessment model had to be chosen and developed. The raters had to be instructed and the rating process had to be carefully thought out. This section will describe general information about all the activities in this study, the choice of scoring the texts with a combination of an analytical model and a multiple trait model, the rationale of choosing certain traits, the instruction of the raters and finally, the rating process.
As explained above the freshmen wrote concept texts and final texts, every concept text and every final text was assessed by two expert raters and two G7 students. The rationale of the rating process can be found in section 3.5.2.2. 20 concept articles were assessed four times, twice by two expert raters and twice by two G7 students. 20 final articles were each assessed four times: two ratings by two expert raters and two ratings by two G7 students. 1 article by Menno Steketee (“Domestic viruses”) was also assessed, also four times, by two expert raters and two G7 students. By assessing an actual text of Menno Steketee, the quality outcomes showed if his style appealed to both the expert raters and the authentic audience. And it also provided information about whether the assessment model for this type of text is valid, as the text of Menno Steketee is the standard the students are supposed to aim for. So if the model is correct, this text will yield high scores on all traits in the assessment model. Thus, a total of 81 different texts in total were assessed. The whole research process in this study is described in table 5.

Table 5: The activities in this study

<table>
<thead>
<tr>
<th>Activities</th>
<th>Quantities</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Students wrote concept texts</td>
<td>20 concept texts</td>
</tr>
<tr>
<td>2. For every text 2 peers and 2 authentic readers wrote a feedback report</td>
<td>40 peer feedback reports 40 feedback reports from the authentic audience</td>
</tr>
<tr>
<td>3. Students wrote final texts, implementing feedback from both sources</td>
<td>20 final texts</td>
</tr>
<tr>
<td>4. The text quality of each concept text (20 texts in total) and each final text (20 texts in total) were evaluated by 2 expert raters and 2 authentic readers using a combination of an analytical assessment model and a multiple trait model. Also the text “Domestic viruses” by Menno Steketee was assessed.</td>
<td>80 ratings for the concept text (20 x 2) and final text (20 x 2) scored by the expert raters 80 ratings for the concept text (20 x 2) and final text (20 x 2) scored by the authentic audience. 4 ratings for the “domestic viruses” text: 2 ratings for the text by Menno Steketee scored by two expert raters and 2 ratings for the text by Menno Steketee scored by two authentic readers.</td>
</tr>
</tbody>
</table>

3.5.2.2 RATIONALE OF CHOOSING AN ANALYTICAL MODEL COMBINED WITH A MULTIPLE TRAIT MODEL

The literature review described the complexity of developing a valid and reliable assessment method. In this thesis I chose to use an analytical model combined with a multiple trait model (see appendix 5 for
the assessment model). The raters had to provide a grade between 1 and 5 for every trait and explain their score. The raters were also requested to give one overall grade for the text; a grade ranging from 1 to 10. This general grade simplifies comparing the texts in overall quality.

I chose an analytical model combined with a multiple trait model, as on the one hand it evaluated general writing ability, for instance spelling, grammar and structure (analytical model), and on the other hand it also focused on audience orientation (multiple trait model). This model then provided diagnostic information about the writing abilities and the audience orientation of the writer, but this information also elucidated the general grade that raters gave on a text. Another benefit of partly choosing an analytical rating scale was that it has been proven to be suitable to train raters, as the scale provides detailed information and is easy to use (Weigle, 2002). The simplicity of being able to use an assessment scale was an important consideration for choosing an assessment method, as young children had to be able to use it, without many problems.

Other valid options to assess the texts would have been primary trait scoring, prominent feature analysis, holistic assessment, comparative methods, or either an analytical model or a multiple trait model separately. Holistic assessment would have been a possibility; a disadvantage of this scale is, however, the lack of detail (Weigle, 2002). I have also decided against using the primary trait scoring or multiple trait scoring separately, as it considers only one trait or more facets of one trait, for instance whether the writer has written a sufficient compare-contrast essay (Weigle, 2002). I could have used this method to assess whether the students have written an audience oriented article, but I also wanted to consider other text features, for instance organization, punctuation and grammar, as these general text features can affect text quality immensely. Prominent feature analysis (Swain, Graves and Morse, 2010) could have been an adequate way to assess these articles for a children’s page, however, discussing all these texts is not practical in a class with G7 students, as it is very time-consuming, although it might be very instructive. The comparative method could have been an option. Even though it is a very reliable method, it does not provide enough detailed insight in the ability of the writers, as it only gives a rank-order of the texts. Using merely an analytical model was also a valid possibility, however, the emphasis of the assignment was audience orientation, and thus, the model should (also) have a strong focus on audience orientation.

Using an analytical scale combined with a multiple trait model also has some pitfalls, for instance that writing should be considered as a whole and not as sub parts added up. Moreover, rating with this model is quite time-consuming, as you have to rate the text per trait. But all things considered, the advantages outweigh the disadvantages.

Using an analytical/multiple trait model in addition to anchor articles would have been a better policy, because it would have made the assessment more reliable. However, letting G7 students assess the articles using benchmark essays, would not have been practical. This method requires a considerable concentration, practice and effort. And it possibly is quite hard for these young students to assess the articles with this method.

3.5.2.3 DEVELOPMENT OF THE ASSESSMENT MODEL

The assessment model (see appendix 5), the mixture of a multiple trait model and an analytical model has been developed in several rounds. The writing assignment and the texts by Menno Steketee have
been studied closely. Reading the writing assignment gave insight into the main aims of this assignment: the students firstly, had to be able to take a young audience into account, secondly, they had to transform difficult scientific texts into a comprehensible text for these readers, thirdly, the article should have the potential to be published on the children’s page of a newspaper and finally, students were asked to imitate the style of Menno Steketee. In sum: a very demanding assignment. Examining the texts by Menno Steketee revealed the features of Menno Steketee’s style: short sentences, short paragraphs, humor and a typical structure. The students had to adhere to these features, to imitate the texts by Menno Steketee. Thus, the assessment model had to include imitation (especially structure and style), transformation of the topic, newspaper suitability and audience orientation.

Second, the texts written by students from previous years were studied for quality. This analysis revealed that grammar, spelling, lay-out and punctuation, can affect quality immensely. These features had to be admitted in the assessment model. These features are also very typical aspects in an analytical model. When examining the written texts, it also became apparent that there are two aspects of audience orientation that can influence whether a text feels appropriate for a certain audience: content and vocabulary. When the writers in this assignment for instance started explaining what graphs are (content) or use a phrase like “a very smart man” instead of researcher, they underestimated the level of the audience. This negatively affected the quality of the text. Thus, audience orientation was included in the assessment model.

Grammar, spelling, lay-out and punctuation were the analytical part of this assessment model. The focus on imitation, audience, newspaper suitability and transformation was the multiple trait part of this assessment model. Even though a typical multiple trait model focuses on only one text ability, for instance persuasion, this model examined these four aspects, as they were all important features for the success of this assignment.

Finally, all these features were fitted into the assessment model. After examination of the model, some traits were less informative than necessary or not specific enough. The trait about audience orientation for instance only questioned if the content was appropriate for the age group, but it did not state specifically what the content for this audience should look like. In the last round, this trait did not only ask to consider if the content was age appropriate, but more accurately asked if the writer used suitable examples and comparisons for this age group. More similar problems were present in the first versions of the assessment model. The model has therefore been revised considerably throughout the development process.

3.5.2.4 TRAITS AND SCORING LEVELS

As said in the previous section, the assessment model had to have various features: imitation, suitability for a newspaper, transformation of the topic, audience awareness and general writing features (lay-out and conventions). In this section I will describe how all the features were represented in the final model.

The imitation of the style of Steketee could be found in several traits in the assessment model: structure, style and imitation. The structure in Steketee’s texts is not very coherent; every new paragraph is usually only indirectly connected to the previous paragraph. To examine whether the writer had used Steketee’s structure, the trait asked whether the text is coherent or not. If the writer wrote a coherent text, he did not adhere to the style of Menno Steketee. Menno Steketee’s style is also very
typical: it uses a lot of humor and is sometimes somewhat twisted. Thus, another trait discussed
whether the style of the text was vivid, attractive and humorous. And finally, the third trait specifically
treated the imitation of Menno Steketee’s texts: did the writer use short sentences and paragraphs, is
humor assimilated in the text and did the writer use his untypical structure? This last trait was meant to
study whether the imitation of the texts by Menno Steketee was successful: a considerable part of this
writing assignment.

The traits title and introduction applied to newspaper suitability and end applies to audience
orientation and partly to Menno Steketee’s style. The title and introduction are of importance for a
newspaper article, as they will determine whether the reader will proceed reading in a newspaper that
has many other articles. The end of an article for children, should finish spectacularly. The trait title
asked whether the title was suitable for a newspaper and was therefore not too long and invited the
reader to read further. The trait introduction discussed whether the reader was triggered to read the
rest of the text. The trait end asked if the last paragraph concluded the text and thus did not peter out;
using an afterpiece was preferable, as also Menno Steketee did this.

More general traits that usually are also included in an analytical scale are lay-out and
conventions. The trait conventions treated the recurrence of spelling, grammar and punctuation errors.
Lay-out was about the beauty of the set-out of the text, how notable the title was and the distinction
between the paragraphs. Both lay-out and conventions were traits in this model, as mistakes could
seriously alter the score on a text. In addition, they were minimum conditions of this course. Writers
that did not fulfill these conditions, received an unsatisfactory mark.

Another trait was the transformation from a scientific text, which meant that the raters
considered whether the students had written an entirely new text for children, and not a reiteration
of the scientific text that functioned as their basic information for the text. This trait had been included
in the assessment model as it was an important part of the writing assignment, as the students had to be
able to convert a difficult text into an easier article.

The last part of the rating model was audience orientation, this was a very important part of the
assessment, as the students were supposed to be able to take their audiences into account. Two
questions were asked about audience orientation. The first trait asked the rater to score the content
appropriateness of the text for this audience and the second trait concerning audience orientation
discussed the suitability of the language for these young children.

The assessment model had to be as specific as possible, to ensure that every rater rated every
trait similarly. Thus every trait had one or more sentences explaining each ‘side’ of the trait. Thus
providing a 1 (very bad) for the introduction is described as: “After reading the introduction I do not
want to read the text”. A 5 (very good) had the following sentence: “After reading the introduction, I
definitively would read the text.” Weigle (2002) notes that raters can understand the words used in the
traits differently, for instance the word ‘attractive’ can be conceived distinctly by different raters. To
prevent this, raters had to be instructed. The instruction of the assessment will be described in section
3.5.2.5.

In this assessment model I had given every trait five scoring levels, as Pollitt (1990, in Weigle)
has stated that raters cannot rate with more than 5 levels reliably. Another reason to use 5 levels was
that also inexperienced raters would be scoring the texts, and Weigle (2002) claims that unskilled
readers are less able than experienced raters in distinguishing between scoring levels.

At the end of the assessment model the raters had to indicate on a scale of 1 to 10 what the
overall quality of the text was. 10 scoring levels were chosen, as in the Netherlands grades range from 1
to 10. This was the easiest scoring measure for all raters, but especially for the grade 7 raters as this
range was already known. This trait was put at the end of the form, as at that point the raters already had made many sub judgements of the text and were therefore able to make a well-considered decision of the quality of the text. The pitfalls and benefits of this method will be described in the discussion.

3.5.2.5 RATING PROCESS

Before the raters could assess the texts, they were instructed on how to assess. The instruction for the G7 students and the expert raters happened at different times. The G7 students were instructed right before they were going to rate the texts, during class. The expert raters were instructed and rated the texts individually at a for them convenient time. All G7 student raters were present at the instruction and 6 out of 7 expert raters were present at this instruction, which I had videotaped, for the one rater that was not able to attend the instruction. She received a file with the instruction.

As said before, the G7 students rated the texts during class and were instructed to ask questions if they required more information. The environment during the rating process was the same for every rater, as they rated in the same classroom. The expert raters assessed the texts in their own homes, but they knew that if they had questions they could contact me at any time. Even though it is preferable to keep the environment of the rating process similar, in this case this was not possible. The experienced raters were all extremely busy and they had to rate approximately 12 texts, thus rating in the same room for several hours, would not have been feasible.

White (1984) advises that two raters assess a text. In this case, not two, but four raters have assessed every text, as I preferred to have two assessments by every rater type. As the expert raters and the student raters will probably rate differently. Moreover, having ratings by more raters provides more information about text quality and enhances reliability.

The 20 concept articles and 20 final texts and a text by Menno Steketee have been assessed by the expert raters and the G7 students. The G7 students all scored 1 concept text and 1 final text. The expert raters rated 5 to 7 concept texts and 5 to 7 final texts. The expert raters thus received an equal mixture of concept texts and final texts to rate. Mixing these texts was to give raters a broad view of the quality of texts. The benefits and pitfalls of the rating process will be discussed thoroughly in the discussion of this thesis.

3.5.2.6 ASSESSMENT INSTRUCTION

The assessment instruction was equal for both the expert raters and the G7 student raters. Before the instruction started, I asked the raters (expert raters and G7 student raters) which features were important to take into account when writing a text for 12/13 year old children. More information about the analysis of the perception of the raters for writing for a young audience can be found in section 3.7.2.1. The instruction itself consisted of two parts: explanation of the assignment and clarification of the assessment model. Concerning the assignment I explained that the students had to translate a scientific article into an article for children of 12/13 years old. And that they had to imitate the texts of Menno Steketee. They received the text “een zilveren randje” (“a silver lining”) by him on paper (see appendix 4 for this article) and I clarified the features of his style using this article. I elucidated that
Steketee uses short sentences, short paragraphs, humor and a typical structure. And that the freshmen students were supposed to also use these features in their articles. For more detailed information on Menno Steketee’s style I refer back to section 3.5.2.4.

Subsequently I instructed the raters about the assessment model per feature. In section 3.5.2.4 I have explained the traits that were used in the assessment model. The traits had to be as specific as possible, and raters had to know what was meant with for instance ‘attractive’. As Weigle (2002) states that raters can understand these labels differently. I have explained the labels I have used in the model: ‘attractive’, ‘vivid’, ‘concluding’, ‘nice’, ‘ugly’ ‘scientific’ ‘suitable’, ‘clear and logical’ and ‘humoristic’. In this explanation I have given specific examples and usually referred back to the example text by Menno Steketee. I have also stressed that the language usage should not be too difficult, but also not too easy. I have given examples of language usage at either end of the spectrum. Saying ‘ten aanzien van’ (regarding) would not be age appropriate for this audience and also too formal for this genre. And for instance by using ‘Meneer Hartley’ (Mr. Hartley), the rater would be highly underestimating the capabilities of the audience. The raters were also explained what was meant with ‘a lot of spelling errors’ and how they were supposed to rate. If there were none or only 1 or 2 small spelling errors or typing errors, the raters were instructed to provide a 5 on the 1 to 5 scale. If there were 10 or more mistakes they were instructed to give a 1 on the scale. If the writer had forgotten to write a title, the raters were instructed to give a 1. All raters were instructed to first read all the texts and then assess all texts, to prevent norm shift.

3.6 OPERATIONALIZATION

In this section the operationalization of the 6 variables will be discussed. The construct validity of each operationalization is described.

The perception of what an appropriate text is for 12/13 year old children was operationalized by asking the expert raters and the G7 students what they thought what was important for a text for this specific audience. They were not leaded in this question. The construct validity of the operationalization of this variable has been accounted for, as 52 raters were asked straight out what features a text for 12/13 year old children were supposed to have. This provides quite a complete image of what aspects are important in an assessment model.

The number of feedback comments have been operationalized by counting all the feedback comments. The construct validity of this operationalization is good, as there is no other way to operationalize the contribution of receiving more feedback.

Implementation of the feedback in this study has been operationalized by the information the students have given in their reflection reports in which they have described exactly which feedback they have implemented in their final texts. The construct validity of this variable has not been guaranteed entirely, as it is uncertain whether the information they have given in their reflection reports corresponds to the actual implementation. This is because the connection between the received feedback and the corresponding revisions have not been studied. This was too time-consuming for the schedule of this thesis.
The number of revisions have been operationalized by counting all the revisions. This was at times difficult, as sometimes changes could be counted as one revision or as two revisions. When was decided how many revisions certain changes were, this was noted and used for following cases. If there was any doubt, a second researcher was consulted. The construct validity of this operationalization is quite good, as a second researcher was consulted in doubt. It, however, would have been better if this second researcher would have also analyzed all the revisions.

The revisions features have been operationalized by analyzing every change made in the text using the taxonomy by Faigley and Witte (1981) or by other categories. These analyses include four categorizations. Firstly, the scope of the revision: surface, word, sentence, paragraph and text as a whole. Secondly, it was analyzed whether text quality on a local level had improved, deteriorated or stayed equal by revising. Thirdly, the influence of revision on the meaning of the text was studied: was the meaning of the text preserved or did the revision change the meaning of the text? These meaning preserving changes or meaning changes are further divided up into: formal changes, meaning preserving for micro changes, meaning altering affects the concept text, meaning altering the summary. Fourthly, the use of different revision operations was analyzed: additions, deletions, substitutions, permutations distributions and consolidations. All these categorizations are explained more extensively in section 3.7.1.6. All the categorizations of the revisions were done by only one researcher, which decreases the construct validity of this operationalization. However, when this researcher was in doubt, a second researcher was consulted.

The features of the revisions by the students was operationalized by studying the feedback reports, the concept texts, the final texts and the ratings. The construct validity of this operationalization cannot be warranted as writing products are analyzed and not the writing process. It is therefore uncertain what actually happened during the revision process.

Text quality was operationalized twice, firstly by a ‘holistic’ score providing an overall score on the text and on sub scores providing ratings for aspects of the text. Overall text quality has been operationalized by asking raters to rate the text on a scale of 1 to 10 at the end of the assessment form. The construct validity of this operationalization cannot be warranted fully, as the raters probably did not provide a unbiased judgement of the text, as they were already leaded by the traits of the assessment model they had already answered. The reason for choosing this method has already been explained somewhat in the end of section 3.5.2.4 Traits and scoring levels and will be further substantiated in the discussion of this thesis. The quality of text features have been operationalized by various traits. The construct validity of this operationalization is probably warranted, as the assessment model has been thought through thoroughly and the raters were trained to use the assessment model. However, as these raters probably have not rated any text before this rating session and are therefore quite inexperienced, it could be possible that the construct validity has suffered slightly. In the discussion the rating process will be discussed extensively.
3.7 ANALYSIS

As mentioned before, the assessment model has a few Likert-scale items; 5 items. The reliability of these items can be measured using Cronbach alpha, this number should be above .7 (De Vellis, 2003 in Pallant, 2007). I have checked the reliability of these items and the Cronbach alpha is 0.75.

In this thesis quantitative analyses and qualitative analyses have been conducted. Both types of research will be explained in this section.

3.7.1 QUANTITATIVE ANALYSES

3.7.1.1 AGREEMENT BETWEEN RATERS ON THE TEXT BY MENNO STEKETEE

The filled-out assessment models of four raters that have rated the text “Domestic viruses” by Menno Stekete are compared. This has been done by studying the differences between the raters in scoring the traits. Those disparities were elucidated by studying the explanations that were given by the raters. This analysis shed light on the validity of the assessment model, as using a text by Menno Stekete was a construct to demonstrate known-groups validity. Known-groups validity signifies that an individual that is known to yield a high score on the construct, also is assessed as having written a text of higher quality. If this text received a high score of the raters on the assessment model, this could mean that the assessment model is valid.

3.7.1.2 AGREEMENT BETWEEN RATERS ON CONCEPT AND FINAL TEXTS

The agreement between the raters on their final scores has also been analyzed. It was not possible to analyze this inter-rater reliability statistically using for instance Cohen’s Kappa, as the raters did not rate enough shared, common texts. The biggest number of shared texts was 7, and sometimes raters did not even rate similar texts. Cohen’s Kappa could therefore not be used. This analysis has now been conducted by examining the differences in ratings between the raters, using descriptive statistics. It was calculated how many percent of the ratings per text differ 0 or 1 points, 2 or 3 points and 4 to 6 points. Thus, if one rater rated a text with a 7, while the other rater scored the text with an 8: there is a disparity of 1 point. It was then calculated how many percent of the ratings differed how many points. This method might not be reliable enough because of the low number of shared texts, however, the inter-rater agreement had to be examined to provide some insight into the reliability of the rating process. As said before, in the discussion the rating process will be discussed extensively.
3.7.1.3 CONTRIBUTION OF THE TRAITS IN THE ASSESSMENT MODEL TO THE FINAL JUDGEMENT

The assessment model was made up of various traits, as also explained in section 3.5.2.4. To assess which traits contributed how strongly to the final score, a regression analysis was conducted. In a regression analysis there is one (or more) independent variable ($X_1$-$X_n$) that can (partly) predict the outcome of the dependent variable ($Y$). The analysis that was used was a commonality analysis. This procedure will be further explicated in 4.5 Contributions of the traits to the final grade.

It is very important to note here, that even though the word ‘prediction’ is used, a regression analysis does not find causal relations between variables. This statistic does not indicate what raters base their final score on. This analysis merely aimed to study the relationship between the traits and the final score, and the statistical contribution of each trait to the final score.

3.7.1.4 ANALYSIS OF IMPROVEMENT

20 concept texts and 20 final texts have been assessed by the grade 7 students and the expert raters. Every text has been assessed twice by the G7 students and twice by the expert raters. This means that one concept text has been assessed four times and the final text has also been assessed four times. In the inter-rater agreement section in the chapter results it was found that the grade 7 raters and the expert raters rated more consistently within their own groups than when both groups were combined. Therefore, all analyses were conducted using the ratings of the grade 7 students and the expert raters separately. For the analysis of improvement for every concept text and every final text the ratings of two expert raters was averaged. The same procedure was adopted for the ratings of the G7 students for the concept texts and the final texts. To examine the difference between the concept text and the final text a Paired Samples t-test was conducted. This analysis was possible for the ratings of the expert raters, however the ratings by the G7 students was not normally distributed, thus the non-parametric Wilcoxon Signed Rank Test was used.

3.7.1.5 THE EFFECTS OF DRAFT QUALITY, FEEDBACK AND REVISION ON TEXT QUALITY

The effects of draft quality, feedback and revision aim to examine the contribution of each of these variables to the overall quality of the text.

As the rating on the concept text is usually a good predictor of the score on the end text, a regression analysis has been conducted. In a regression analysis there is one (or more) independent variable ($X_1$-$X_n$) that can (partly) predict the outcome of the dependent variable ($Y$). Thus, in this case the regression analysis studied how the fixed independent variable score on concept text, affected the dependent variable score on final text. It should be noted here, that self-evidently not only the score on the concept text can influence the final score, other variables can also affect this score, for instance revision or the other factors described in the literature review: writing self-regulation and self-efficacy,
writing approaches, writing apprehension and writing beliefs. These factors are not analyzed in this thesis, however, it should be borne in mind that these factors can affect text quality. This analysis only aimed to study what the unique contribution of the score on the concept text is on the rating of the final text.

To examine the influence of feedback on end judgement, it was important to also investigate revision, as they are narrowly related. For a writer to implement feedback properly, he has to know how to revise. The feedback was analyzed by studying how many feedback points the writers had received. Thus, a feedback provider can provide feedback on the introduction, the spelling error in the second paragraph and on the usage of a word that does not seem appropriate for the audience in paragraph 3. In this case the writer received 3 feedback comments. Not only the number of feedback points have been studied, also the total number of feedback points that have been implemented in the final texts (implemented feedback), according to the writer’s reflection reports was used in the analysis. A writer who received 3 feedback points (number of feedback) could have chosen to only implement 1 feedback point (implemented feedback). The influence of each of these variables on text quality was analyzed using regression analyses. Important to note here is that the relationship between feedback and text quality is in no way straightforward. Raters can decide not to do anything with the feedback or they can decide to use the feedback, but then revision mediates between feedback and text quality. The reader should therefore bear in mind that these analyses are extremely tentative.

In addition, the influence of the number of feedback comments on the number of revisions in the final text is analyzed using a regression analysis.

Also the number of revisions are counted and taken as a measure: the writer can for instance change a spelling error, the sequence of a sentence, punctuation in another sentence and alter the title of the text. This writer then has made 4 revisions. The influence of revising more on text quality has also been analyzed using regression analyses.

3.7.1.6 ANALYSIS OF THE REVISIONS OF FRESHMEN STUDENTS

Also the revisions of the freshmen students have been analyzed. For this analysis the revisions were divided up into categories partly based on the taxonomy by Faigley and Witte (1981) (see section 2.2.3.1 Types of revision in the literature review for more information about their taxonomy) and partly by other categories as also explained earlier. Faigley and Witte make a distinction between revisions that can change the meaning of the text and those that can preserve the meaning of the text. In this thesis four distinctions have been made for these meaning changes and meaning preserving changes. 1) formal changes, which were all the changes that do not actually change the text at heart, but just at the surface: punctuation changes, abbreviations, lay-out, spelling and etcetera. 2) Meaning preserving for micro changes: which signified that the writer changed something on micro level (word-level) that did not affect the meaning of the text. 3) Meaning altering affects concept text: the revision changed the meaning of the concept text. And finally, 4) Meaning altering the summary: changes in the text change the summary of a text if a summary was made: very rigorous changes. The influence of each of these types of meaning changes was analyzed using descriptive statistics. It was analyzed how frequently every type occurred in the revisions made by the freshmen students.
Another part of the taxonomy by Faigley and Witte (1981) that was used in these analyses were the operations that could be used to revise a text: *additions, deletions, substitutions, permutations, distributions* and *consolidations*. These operations have been used exactly as in the taxonomy of Faigley and Witte. For an explanation of these operations I would like to refer the reader back to the literature review: 2.2.3.1 *Types of revision*. The occurrence of these operations has also been analyzed using percentages.

The scope of the revision has also been studied. This analysis is not based on the taxonomy by Faigley and Witte. *Scope* had five categories: *surface* (changing spelling, tense, punctuation and etcetera), *word, sentence, paragraph* and *text as a whole*. The occurrence of these categories was also analyzed using descriptive statistics, thus the occurrence of these categories.

The last analysis for revisions was to examine whether the local revisions (for instance changing one word) improved the text, deteriorated the text or kept the text similar in quality. These categories were also analyzed using percentages.

3.7.2 QUALITATIVE ANALYSES

3.7.2.1 PERCEPTION OF A SUITABLE TEXT FOR 12/13 YEAR OLD CHILDREN

The perception of the expert raters and the authentic raters what the features were of an appropriate text for 12/13 year old children. No analysis has been used, merely the answers of the raters on this question will be described in the result section of this thesis.

3.7.2.2 STUDY OF REVISIONS

Another qualitative analysis that was conducted was a study of the revisions of two writers, as the writers of these texts either successfully or un成功fully revised their texts according to the ratings by the raters. Both texts are examined using the concept text, the final text, the feedback, the revisions and the assessments. All aspects were described and it was hypothesized how the strong deterioration and improvement could have occurred.

The next chapter will describe the results of the analyses, which will lead to the answers on the research questions.
4. RESULTS

4.1 INTRODUCTION

This chapter is structured in four parts: firstly the reliability and validity of the assessment model is discussed, secondly, the contributions of all the traits to the final judgements of the texts is elucidated, thirdly, the contribution of feedback and revision to text quality is described and finally the revisions of the students is explained more thoroughly.

In this chapter the reliability and validity of the assessment model is discussed in various ways. Firstly, the raters were asked what their perception was of a suitable text for children, secondly, the inter-rater agreement of the text written by Menno Steketee is described, as this can demonstrate whether the assessment model is valid and, finally, the inter-rater agreement of all the texts written by the freshmen students is discussed.

The second part of this chapter describes the contribution of all the traits to the final judgements. This analysis provides some insight into which traits have contributed most to the final judgement of the texts. The judgements of the traits are divided up into the ratings by the grade 7 students and the expert raters.

The third part of this chapter describes the contribution of feedback and revision to the final judgement of the text. It was for instance analyzed whether revising more has a positive contribution to the quality of the final text.

The fourth and last part of this chapter focuses on revision. Revision plays a significant role in text quality, as revising globally or locally can have positive effects on text quality (Faigley & Witte, 1981). In this part the revision practices of the students are described using descriptive statistics. And to illustrate the revision strategies of the students, the practices of two students are described more extensively.

4.2 PERCEPTION OF A SUITABLE TEXT FOR CHILDREN

As described in the introduction of this chapter, the validity and the reliability of the assessment model will be discussed thoroughly. However, the validity of this assessment model also depends on whether the focal points used in this assessment model were actually important for an assessment model for children. In other words, if the aspects in the assessment model, for instance style and structure, were actually aspects that are considered significant for a popular scientific text for children. To examine this, both the grade 7 students and the expert raters were asked about their perceptions of what a suitable text is for children of 12/13 year old, to enquire if their focal points fitted with the aspects used in the assessment model. Asking both types of raters about their opinions on texts, also provided some insight into how well these raters were able to distinguish important features in a text. Thus, whether they knew which aspects in a text are important for children. This enquiry was held before they had to rate all the texts.

The grade 7 students and the expert raters mentioned some similar aspects, but also made their own assumptions. Both groups indicated that a text for this age group should not be too simple, but also
not too difficult. The expert raters stated specifically that writers should not underestimate their readers. They also both mentioned that there should be a logical structure in the text; the experts stated that writers should use connectives.

Expert raters also mentioned that you should explain a lot, be playful, throw in a few jokes and that readers should be able to identify themselves with the text. In accordance with that, the expert raters claimed that the topic should be geared to the reader’s perception of their environment. And finally, they said that a writer should use somewhat shorter sentences.

The grade 7 students mentioned that the sentences, paragraphs and the text had to flow well, the topic of the text should be enjoyable and there should be a good introduction. The class did not agree on whether the text should have an illustrating picture. One student said that he liked having a picture with a text, the other said that a picture prevents him from making his own image of a text.

These short interviews indicated that both the expert raters and the G7 raters recognize the importance of audience orientation; and specifically attuning to the proper level of difficulty appropriate for 12/13 year old children. Structure was also mentioned by both types of raters. A notable difference between the perceptions of both raters was that the expert raters were more specific: for a good structure, expert raters claimed that connectives should be used, and to reach your audience you should throw in jokes and be playful. The grade 7 students merely claimed that the text should ‘flow’, how to attain ‘flow’ was not specified by these students. They also merely stated that the topic of the text should be enjoyable and the text should not be too hard or too difficult, but they did not provide any other indications of how to adapt the text to 12/13 year old children. The expert raters described more specifically how to achieve audience orientation: for instance attuning to the reader’s perception of their environment, being playful, using shorter sentences and explaining a lot. It seems that expert raters are better aware of which strategies can be used to achieve audience orientation. The G7 student raters seemed to know what they want, but not how to attain it. These differences suggest that the G7 students are less able to distinguish important features in the text.

This enquiry suggested that the focal points used in the assessment model match with the perceptions of both types of raters, as all the aspects mentioned by the raters were also represented in the assessment model.

4.3 AGREEMENT ON THE MENNO STEKETEE TEXT

As said before, the freshmen students had to imitate the texts of Menno Steketee, an author with an unusual style characterized by humor, short sentences, short paragraphs and a typical structure. All raters were informed of the assignment of the students; they knew the freshmen had to mimic the features mentioned above. The assessment model has been developed based on the characteristics of the texts of Menno Steketee, as this was the standard the freshmen had to reach with their texts. If this assessment model was correct, a text of Menno Steketee would yield a high rating. Thus, four raters have also assessed a text by Menno Steketee, titled Tamme virussen (“Domestic viruses”, see appendix 4 for this text), to see if it would actually get a good rating. If the raters provide a high rating for this text, this also suggests that the raters are capable of recognizing good quality. Using a text by Menno Steketee is a construct to demonstrate known-groups validity. Known-groups validity signifies that an
individual that is known to yield a high score on the construct, also is assessed as having written a text of higher quality.

The raters all agreed that the text by Menno Steketee was of good quality; all four raters provided a 9 on a 10-point scale as a final score. One grade 7 student indicated that the text resembled the exemplary text they read by Menno Steketee: *Een zilveren randje*, (“A silver lining”, see appendix 4 for this text). An expert rater stated that this text was a “good imitation.”

All raters have scored this text with a 9, and this suggests that the assessment model is a good model to assess texts written for this particular assignment. However, the four raters did not rate all bands with fives, thereby indicating that there is room for improvement. Table 6 shows how all four raters have assessed the text on band scores using the assessment model.

### Table 6: Band scores and the total score on the text by Menno Steketee, by two expert raters and two grade 7 student raters

<table>
<thead>
<tr>
<th></th>
<th>expert rater 1</th>
<th>expert rater 2</th>
<th>grade 7 student 1</th>
<th>grade 7 student 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Introduction</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>End</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Audience content</td>
<td>5</td>
<td>5</td>
<td>- (no answer)</td>
<td>5</td>
</tr>
<tr>
<td>Audience language</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Structure</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>Style</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Spelling and grammar</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Lay-out</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Transformation science – children’s article</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Imitation</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>End judgement</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
</tbody>
</table>

The bold font indicates a rating lower than 5 on the band scores.

As the table shows, two expert raters and one grade 7 student rater have assessed the text of Menno Steketee with excellent band scores. Only grade 7 student two was less enthusiastic about the text, if you examine the band scores and not the final score.

Expert rater one rated nine of the eleven bands with a 5: lay-out with a 4 and structure with a 3. The other expert rater has assessed this text quite similarly; also nine out of eleven bands with a 5: lay-out with a 4 and introduction with a 3. They both agreed that the lay-out could be improved. These scores can be accounted for as I have retyped the text verbatim, without imitating the existing lay-out, to lower suspicions that this text could be written by Menno Steketee. Expert rater two also thought that the introduction could be improved; her 3 can also be easily explicated: Menno Steketee has an introduction of 1 sentence in his text; this is very short, and possibly has caused the low rating. The comment of the rater on her 3 on the introduction subscribed this hypothesis: “Not very clear, only one sentence”.
Grade 7 student two almost agreed fully with the expert raters when the band scores were examined, he also rated 9 out of 11 bands with a 5. Only the score of 1 on imitation is conspicuous. When examining his assessment forms this appeared to be an error, as it seemed as if this rater had not understood this trait of the assessment model. He commented namely that the text at issue did not have the same topic as the exemplary text of Menno Steketee (“A silver lining”). On the form of the other text he rated provided the same comment. This indicated that he did not fully understand what this item meant. According to this rater also the end and the language could have been better. He, however, did not provide any critical comments on either of the items: the end was “logical” and the text “did not contain difficult words”. Thus, it was hard to specify what exactly the rater thought what could be improved.

The other grade 7 student rated 4 out of the 11 bands with a 5. The title has been rated with a 4; introduction with a 2; end with a 4; audience language with a 3; structure with a 2 and spelling and grammar with a 4. This grade 7 rater provided introduction with a low band score, just as expert rater two; commenting that it is “quite short” and that it did not “relate to the title.” She also assessed the structure as poor, just as expert rater two. She indicated that the topics were not really connected, just a little. This is actually a typical Menno Steketee structure, this was also explained in the rater training. Apparently the explication of this trait was not done clearly enough. The title did not receive a critical remark; she stated that she was “curious what domestic viruses are”. The end of the text could be improved according to this rater by adding “something”. She provided audience language with a 3, as it was not entirely unsuitable for children of her age, but also not completely appropriate for the audience. Spelling and grammar also received a 4, even though the rater did not find any mistakes in the text. The low scores as compared to the other three raters and the 4 on spelling and grammar, even though she did not find any mistakes, could indicate that this rater scored very strictly. This suggestion was endorsed if you examine the bands of the other text she has assessed. She provided even lower band scores on this other text: 8 of the band scores are 3 or under, though she still gave a 7 as a total score on the text.

All raters were however enthusiastic about the text, as they all rated it with a 9. This suggests that the raters were able to see good quality and that the assessment model was able to distinguish a good quality text from a text of poorer quality. It should be noted here, however, that the band score on structure received a low score from both a grade 7 student and an expert rater; apparently this trait description was confusing for the raters. This trait stated that the connections between the paragraphs are logical and clear. However, I did explain that in the texts by Menno Steketee the paragraphs were only indirectly linked to each other. The raters must have focused merely on the traits, causing a low score on structure. Thus, the item structure did not capture the typical structure of Menno Steketee’s style.

4.4 INTER-RATER AGREEMENT

As said in the method section of this thesis were the 20 concept text and 20 final drafts of the texts evaluated four times; every text has been assessed by two expert raters and two grade 7 students. They were asked to provide an overall rating for the text on a scale of 1 to 10.

In this paragraph I will report on the inter-rater agreement between the expert raters as a group, between the grade 7 students as a group and between the expert raters and the grade 7 students.
together. Inter-rater agreement is highly important, as agreement between the raters affects the reliability of the ratings.

As also discussed in the method section, the inter-rater agreement could not be analyzed using statistics, as the raters had rated too little shared texts. Thus, the inter-rater agreement on the assessments of both the concept texts and the final drafts of the texts were analyzed using descriptive statistics. The analysis compared ‘general’ raters, thus the expert raters and the grade 7 raters, passing over the individual rater. The discrepancies between the ratings were analyzed by counting how frequently all discrepancies between ratings of texts occur. Thus, for instance if one rater rated a text with, for instance, a 7 (on a 10-point scale) and another rater with a 5, there was a difference in the rating of 2 points. It was calculated how frequently each discrepancy occurred for every group. In this section, if a difference of 4 or more occurred in the ratings, these discrepancies will be explained.

4.4.1 CONCEPT TEXTS EXPERT RATERS

7 expert raters have assessed the 20 concept texts; all concept texts have been assessed twice by the expert raters. They had to indicate the quality of the texts on a 10-point scale. A discrepancy was accepted when the two expert raters differ 0 or 1 in their rating. A discrepancy of 2 or more was not desirable. 75% of the ratings differed 0 or 1 point from each other. The remaining 25% consisted of a discrepancy of 2 or 3 (15%) and a difference of 4 or 5 (10%). Differing 0 or 1 point in 75% of the ratings is a respectable agreement.

The considerable difference of 4 or more was due to raters rating extremely. One rater claimed one text to be of poor quality, rating it with a 2, stating that this text was too difficult for these students. While the other expert rater rated this text with a 7. Also the grade 7 students rated this text with a 7. This final score of 2 seems to be an outlier. It might be possible that this rater had difficulties estimating the level of the audience.

The other considerable discrepancy was also due to an extreme rating. This rater had rated the text with a 9, while the other rater had provided this text with a 5. They did not seem to agree about the rating of this text. The 5 has been provided for this text, as the rater claimed that the title was too uninspired, that the text had a bad introduction and a poor end, the structure was mediocre and there was no humor. This 5 seemed to be a corollary of all the poor features of the text. The rater who provided this text with a 9 has made many comments, claiming that the sentences did not always ‘flow’, some words were too childlike and she stated that the text was not that funny. After critiquing the text in this manner and then providing the text with an overall grade of a 9 seems a bit odd.

The first considerable discrepancy might be due to a rater that did not seem to have an understanding of what a suitable difficulty level is for these 12/13 year old children. The other rater provided the text with an overall grade of a 9, even though her separate ratings were less positive. This high rating could be due to the rater-bias personal comparison, in which a rater has the tendency to rate very strictly or lenient.
40 grade 7 students have assessed the 20 concept texts, every concept text has been assessed by two grade 7 students, thus 40 ratings have been compared. The agreement between the assessors was good, when the raters only differ 0 or 1 in their opinion of the quality of the text. A rating difference of 2 or more was not preferable. 50% of the ratings differed 0 or 1 point from each other. The remaining 50% consisted of a discrepancy of 2 or 3 (35%) and a difference of 4 (15%). 50% agreement for 0 or 1, was a less than moderate agreement. The expert raters agreed more on the ratings of the concept texts (75% differ 0 or 1 point from each other) than the grade 7 raters (50% differ 0 or 1 point). This could suggest that the expert raters have rated more consistently than the grade 7 students.

As said above, a difference of 4 points occurred in 3 ratings. The first considerable difference in ratings was due to a grade 7 rater who rated the text with a 4, as he also had to rate the text by Menno Steketee, which was of better quality than the text by the freshmen student. As this student compared the quality of both texts, norm shift occurred and the text by the student seemed to be of very poor quality compared to the text written by Menno Steketee.

Another text was rated very differently by one grade 7 student, as this rater did not feel that the topic of the text, death and diseases, was suitable for children and he therefore evaluated the text with a 4. The other raters (grade 7 students and expert raters) did not seem to have a problem with the topic and rated it on average with a 6.3.

The third and last big difference between the ratings was due to a rater who provided the text with a 7, while the other rater rated this text with a 3. The rater who provided this text with a 7, had a tendency to give high grades: an expert bias that is called personal comparison. Even though he provided this text with a 7, he was very critical in his assessment model and stated that the text was too childlike and that the structure was not that good. In another text he had to rate, he was also very critical, but he still provided the text with a high grade. It seemed as if this rater has the tendency to rate highly.

To calculate the inter-rater agreement between the expert raters and the grade 7 students, their ratings (N=80) were compared. Expert one was compared with grade 7 student one and grade 7 student two. Expert two was also equated to grade 7 student one and grade 7 student two. In total, there were 80 comparisons. 55% of these comparisons differ 0 or 1 from each other. The remaining 45% consists of 37.5% with a discrepancy of 2 or 3. 7.5% of the cases differed 4 or more. A difference of 0 or 1 in 55% of the cases is a less than moderate agreement.

This analysis suggested that the expert raters and the grade 7 raters taken together rated less consistently, than when the raters were compared as subgroups. The judgements by the expert raters and grade 7 students for the concept texts has been represented in figure 3. This figure shows boxplots of the ratings by the two expert raters and two grade 7 student raters for every concept text. Most boxplots are quite long, which means that there were enormous differences between the ratings on the text.
texts. A few boxplots are not that tall, which means that the raters agreed on the quality of the text. Only text 10 has been rated similarly by all four raters.

![Figure 3: Ratings on the concept texts by both types of raters](image)

4.4.4 FINAL TEXTS EXPERT Raters

7 expert raters have also assessed the final texts. Expert rater one has been equated with expert rater two. 19 final text ratings could be compared, as one expert rater had neglected to fill in one rating of one text. The expert raters’ assessments agreement was moderate: they differed 0 or 1 in 53% of all assessments of the final texts. For 42% of the texts they differed 2 or 3, in 5% of the texts they differ 4 or more (1 rating).

One text has been rated with a 2 and an 8; an enormous discrepancy of 6 points. This difference was due to a rater who did not deem this text as appropriate for this audience, as it lacked humor and
examples according to this rater. This rating seemed to be an outlier, and was made by the same rater who also provided the concept text, described above, with a 2. The impression of one rater rating differently was strengthened by the ratings made by the grade 7 students. They have provided the text with a 7 and an 8.

40 grade 7 students assessed 20 final texts, and also here, the score of grade 7 student one has been compared to the rating of grade 7 student two. The grade 7 students were also in moderate agreement, they differed 0 or 1 in 55% of the ratings. They disagreed on 2 or 3 in 30% of the cases and 4 or more in 15% of the scores. Also for the final texts the grade 7 students seemed to differ more than the expert raters, as 15% of the scores differed 4 or more, while this was 5% for the expert raters.

There were three ratings that differed 4 or more. The first ratings differed, as these raters simply disagreed. One rater considered this text to be of poor quality as certain simple information was explained, even though it did not need explaining. This rater has rated this text with a 3. The other grade 7 student was very positive about this text, he enjoyed the topic, he thought it was a clear text and he said it was suitable for his age and rated this text with a 9.

The second discrepancy was also due to simple disagreement. One rater thought the text was suitable for his age and he stated that the text was clear and funny, and provided the text with a 9. The other rater claimed that the writer had highly underestimated the audience: the text was too childlike and had too many explanations.

And the last big difference in ratings was a text that was provided with a 9 and a 4. The expert raters were also not that positive about the text; they rated the text with a 3 and a 6. It seemed as if the 9 is an outlier, which was also most probably due to disagreement. This rater was very positive about the text, thought the text was appropriate for this age group and stated that the text is clear.

All these ratings could not be explained by rater biases as norm shirt or personal comparison, but were most probably due to simple disagreement between the raters about what is considered to be a text of good quality.

38 text ratings of both groups, grade 7 students and expert raters, have been compared. In this analysis not 40 texts have been considered, as one expert rater did not fill in a rating for one final text. Thus, it was not possible to compare a rating of one text of one expert rater, with grade 7 student one and grade 7 student two. The grade 7 students and the expert raters were also in moderate agreement when rating the final texts: they differed 0 or 1 in 53% of the text ratings. They disagreed 2 or 3 in about 33% of the final scores. For the remaining 14% the raters disagreed 4 points or more.

Above it was described for the concept texts that when the expert raters and the grade 7 students were compared in one group for their inter-rater agreement, they seemed to rate less
consistently than when the groups were compared separately. Also for the final text, as the analysis above shows, the grade 7 students and the expert raters rated more consistently when they were compared within their own group. This suggested that the raters have different viewpoints of a suitable text for 12/13 year old children. In figure 4 an overview of the ratings by the expert raters and grade 7 students for every final text can be found. Similar to the distributions of the ratings of the concept text, also the final texts have been rated very differently. Some boxplots are extremely tall, for instance for the ratings on the final texts 5 and 19, which means that the ratings on these texts differed extremely. One rater has rated text 5, for instance, with a 3, while another rater at the other end of the boxplot, has provided this text with a 9. Text 11 has been rated similarly for all raters, namely with a 7.

![Figure 4: Ratings of the final texts by both types of raters](image)

In table 7 an overview has been given of the agreement between grade 7 students, the discrepancies in ratings between the expert raters and the agreement between the grade 7 students and the expert raters. Data have been provided for the concept texts and the final texts.
Table 7: Inter-rater agreement of the ratings on the concept and final texts in the expert raters group and in the grade 7 students group and also between the expert raters and the grade 7 students, in percentages.

<table>
<thead>
<tr>
<th></th>
<th>Concept E</th>
<th>Concept G7</th>
<th>Concept E &amp; G7</th>
<th>Final E</th>
<th>Final G7</th>
<th>Final E &amp; G7</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference of 0 or 1</td>
<td>75%</td>
<td>50%</td>
<td>55%</td>
<td>53%</td>
<td>55%</td>
<td>53%</td>
</tr>
<tr>
<td>Difference of 2 or 3</td>
<td>15%</td>
<td>35%</td>
<td>37.5%</td>
<td>42%</td>
<td>30%</td>
<td>33%</td>
</tr>
<tr>
<td>Difference of 4 or more</td>
<td>10%</td>
<td>15%</td>
<td>7.5%</td>
<td>5%</td>
<td>15%</td>
<td>14%</td>
</tr>
<tr>
<td>N</td>
<td>40</td>
<td>40</td>
<td>80</td>
<td>19</td>
<td>20</td>
<td>78</td>
</tr>
</tbody>
</table>

This table shows that there was mostly a moderate agreement within the groups, but also between the groups, with on average a difference of 0 or 1 for 57% of the ratings. Only the concept texts were assessed quite evenly by the expert raters, with an inter-rater agreement of 75%, in which they differed 0 or 1 in their ratings.

The analyses of the inter-rater agreement of the concept texts and the final texts indicated that the expert raters and the grade 7 students rated more consistently when they were regarded as separate groups. When they were considered as one group, they disagreed more frequently. Thus, in the remainder of this thesis will the ratings of the expert raters and grade 7 students therefore be analyzed separately, to increase the reliability of the assessment.

4.5 CONTRIBUTIONS OF THE TRAITS TO THE FINAL GRADE

The texts have been examined by the grade 7 raters and the expert raters. All the raters had to indicate their judgement of the text using an assessment model with various traits. In this section the influence of these traits on the final judgement of the text will be discussed. Some traits in the assessment model were connected to each other, as raters could use the same criteria for traits that were somewhat similar. This could be the case for for instance both traits that focused on audience orientation; the content of the text and the language of the text and whether these were fitted to the audience. These traits partly overlapped and therefore could be correlated to each other. The correlations between the traits were therefore first examined, and if these connections were also meaningful, those traits were combined. Only the correlations that were 0.2 or higher will be considered. After describing the considerations to combine certain traits, the contribution of the traits on the final judgement will be elucidated.
4.5.1 CORRELATIONS BETWEEN THE TRAITS IN THE ASSESSMENT MODELS

4.5.1.1 AUDIENCE CONTENT AND AUDIENCE LANGUAGE

The traits *audience content*, *audience language* and *transformation to an article* were correlated. These correlations by the expert raters and grade 7 students separately were between 0.3 and 0.6. The strength of these correlations is medium to large.

*Transformation to an article* was correlated with *audience content* and *audience language* statistically for both the expert raters and the grade 7 students. For the expert raters the correlation between *audience content* and *transformation to an article* was 0.5, a correlation of medium strength. *Audience language* and *transformation to an article* were correlated with 0.3, also a correlation of medium strength. For the grade 7 students the correlation of *audience content* and *transformation to an article* was 0.5 (medium strength). And *audience language* and *transformation to an article* was 0.3 (medium strength). These traits were not only connected statistically, but were also connected in meaning. These traits were therefore joined.

To examine the reliability of this combination; thus whether these variables were linked, a reliability analysis has been conducted. Cronbach’s alpha for this combination, when the texts were rated by the experts, Cronbach’s alpha is 0.71. When the texts were rated by the grade 7 students Cronbach’s alpha is 0.69. The consistency of the combination of these three variables is questionable to acceptable.

Even though *imitation* was correlated with *audience content*, *audience language* and *transformation to an article*, they were not connected in meaning. Concerning meaning, *style* could be connected with *audience content*, *audience language* and *transformation to an article*, however, *style* can be better linked with *imitation*, as will be described later in this section.

4.5.1.2 TITLE, INTRODUCTION AND END

*Title*, *introduction* and *end* were not combined in this analysis. These traits were somewhat correlated, but they were not linked in meaning. Moreover, a reliability analysis revealed that these traits did not cohere: expert raters Cronbach’s alpha= 0.63 and grade 7 students: 0.35. These Cronbach’s alphas are questionable to unacceptable. The traits *title*, *introduction* and *end* also correlated with other traits in the assessment model, for instance *style*, *audience content* and *imitation*. However, even though these correlations are explicable, as *style*, *audience content* and *imitation* will contribute to *title*, *introduction* and/or *end*, these traits were not joined, as these traits were not meaningfully linked to each other.
4.5.1.3 STYLE AND IMITATION

As already alluded to above, style and imitation correlated also with almost all variables described above. However, style and imitation were best connected together, as these traits were closely linked in meaning. The style of Menno Steketee was a large part of a proper imitation of a text by Menno Steketee. The trait style could therefore be seen as a part of the trait imitation. Both types of raters have seen this connection, as these variables are highly correlated: \( r=0.6 \) \( (p < 0.01) \) for the grade 7 raters and \( r=0.7 \) \( (p < 0.01) \) for the expert raters. These variables were therefore joined into the new variable: imitation. The corresponding Cronbach’s alpha for the expert raters was 0.82 and for the grade 7 raters: 0.62. The consistency of all these combinations is questionable to good.

4.5.1.4 STRUCTURE

Structure for the grade 7 students correlated with lay-out \( (r=0.4, p < 0.01) \) and introduction \( (r=0.2, p < 0.01) \). For the expert raters structure was only correlated with lay-out \( (r=0.4, p < 0.01) \). Only the link between lay-out and structure was taken into account, as this correlation was joined for both the expert raters and the grade 7 raters. The connection between structure and lay-out was understandable, as the assessment model asked the raters to take the blank lines between the paragraphs into account. However, structure could not entirely be connected to lay-out, as structure should be more focused on content than on form. The trait structure has therefore not been connected to lay-out or to any other variable.

4.5.1.5 SPELLING ERRORS AND PUNCTUATION AND LAY-OUT

Spelling errors and punctuation did not correlate to any variables when the judgements on the traits by the expert raters have been examined. When the grade 7 students rated the quality of the texts, there was a small correlation between spelling errors and punctuation and lay-out \( (r=0.2, p < 0.01) \). Lay-out correlated with structure, as said above, and also with audience language for the grade 7 students \( (r=0.5, p < 0.01) \), which seems quite odd, as these traits were not even slightly similar in meaning. Lay-out also correlated with title \( (r=0.3, p < 0.01) \) for the grade 7 students, which is more explicable, as the size or fond of the title can contribute positively or negatively to the lay-out. This, however, could mean that the grade 7 students did not adhere to the assessment model closely.

Even though lay-out and spelling both focused on lower order concerns they did not have a connection in meaning. The consistency of their combination was also very low when rated by the expert raters and the grade 7 raters, Cronbach’s alpha was between: 0.2 and 0.3. These variables were therefore not joined.
When all variables were joined there were two new variables: *audience* and *imitation*. *Title, introduction, end, structure, lay-out* and *spelling errors and punctuation* remained unconnected. *Audience, title, introduction, end, imitation* and *structure* were the higher order concerns and *lay-out* and *spelling errors and punctuation* the lower order concerns.

### 4.5.2 CONTRIBUTIONS OF THE TRAITS TO THE FINAL JUDGEMENT

#### 4.5.2.1 EXPERT RATERS

To examine which aspects contributed to the final judgement the most, regression analyses have been conducted for the ratings of the expert raters. The model was significant at the 0.00 level for all analyses, which means that the traits made a significant contribution to the final judgement. The quality differences in all traits predicted 77 percent of the quality differences of the final judgement (Adjusted $R^2=0.77$). This meant that the judgement of the text could be attributed to the assessment model for 77 percent. 23 percent remained hidden.

To be able to study which traits contributed most to the explained variance, a commonality analysis (following De Glopper, 1983) has been conducted for all separate traits. First a model was estimated with all the traits; this was the total explained variance. To analyze the unique contribution of one trait, for instance *lay-out*, to the total explained variance (in the above case 0.77) a new model was estimated in which this trait was removed. The difference between the total explained variance and the variance in which the separate trait was removed, was the uniquely explained variance of that trait. The common contributions were also analyzed. This means that the combination of the traits, for instance *audience, imitation* or *higher order concerns*, could also contribute to the explained variance together, as a remainder, when the unique variance of the traits within the combination were subtracted.

In table 8 all these unique contributions of the separate traits and the common contributions are represented.

**Table 8:** The prediction of each text aspects to the final judgement of the text, when the texts were rated by the expert raters

<table>
<thead>
<tr>
<th>Audience content</th>
<th>Audience language</th>
<th>Transformation to an article</th>
<th>Style</th>
<th>Imitation</th>
<th>Title</th>
<th>End</th>
<th>Introduction</th>
<th>Structure</th>
<th>Lay-out</th>
<th>Spelling errors and punctuation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj. $R^2=0.006$</td>
<td>Adj. $R^2=0.007$</td>
<td>Adj. $R^2=0.02$</td>
<td>Adj. $R^2=0.006$</td>
<td>Adj. $R^2=0.002$</td>
<td>Adj. $R^2=0.04$</td>
<td>Adj. $R^2=0$</td>
<td>Adj. $R^2=0$</td>
<td>Adj. $R^2=0.001$</td>
<td>Adjusted $R^2=0$</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Audience</th>
<th>Imitation</th>
<th>Higher order concerns</th>
<th>Lower order concerns</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted $R^2=0.04$</td>
<td>Adjusted $R^2=0.02$</td>
<td>Adjusted $R^2=0.57$</td>
<td>Adjusted $R^2=0$</td>
</tr>
</tbody>
</table>
Common
Adjusted $R^2 = 0.04$

All these aspects together: Adjusted $R^2 = 0.77$

Table 8 demonstrated that the contribution of the separate traits was quite low: the end uniquely explained 4% of the variance. *Style* also contributed to the variance, with 2%.

From the combinations, *audience* communally contributed most to the total explained variance with 4 percent. The communal contribution of the higher order concerns - when all the separate traits and the combinations were subtracted - was 57 percent. Apparently the traits that were part of the higher order concerns were hard to distinguish and overlapped when explaining the unique variance.

*Layout* and *spelling and punctuation* hardly explained any variance. *Structure* and *introduction* contributed nothing to the final judgement of the text. *Layout* uniquely explained 0.1% of the variance and *spelling and punctuation* did not contribute to the final judgement. The lower order concerns communally, in which *layout* and *spelling and punctuation* have been subtracted, explained nothing of the total variance.

When all variables and combinations have been subtracted from the total explained variance, *Common*, remains. *Common* only contributed 4 percent of the total explained variance. This means that the lower and higher order concerns could be distinguished and hardly overlapped.

These regression analyses showed that the communal higher order concerns, and the combination, contributed most to the final judgement of the text. The lower order traits did not predict the final judgement of the text.

### 4.5.2.2 Grade 7 Students

The focus of the grade 7 students has also been examined. To study which traits predicted the final judgement the most, regression analyses have been conducted, also these models were significant at the 0.00 level. The quality differences in all traits predicted 59% of the quality differences in the final judgement (Adjusted $R^2=0.59$). The influence of the other separate traits has been represented in table 9 below.

**Table 9:** The prediction of each text aspect to the final judgement of the text, when the texts are rated by the grade 7 raters

<table>
<thead>
<tr>
<th>Audience content</th>
<th>Audience language</th>
<th>Transformation to an article</th>
<th>Style</th>
<th>Imitation</th>
<th>Title</th>
<th>End</th>
<th>Introduction</th>
<th>Structure</th>
<th>Layout</th>
<th>Spelling and punctuation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adj. $R^2=0.001$</td>
<td>Adj. $R^2=0.02$</td>
<td>Adj. $R^2=0$</td>
<td>Adj. $R^2=0.04$</td>
<td>Adj. $R^2=0$</td>
<td>Adj. $R^2=0$</td>
<td>Adj. $R^2=0.02$</td>
<td>Adj. $R^2=0.03$</td>
<td>Adj. $R^2=0.005$</td>
<td>Adjusted $R^2=0.03$</td>
<td>Adjusted $R^2=0.02$</td>
</tr>
</tbody>
</table>

Audience
*Adjusted $R^2=0.04$*

Imitation
*Adjusted $R^2=0$*
Table 9 shows that the contributions of the separate traits was quite low; the highest contribution was four percent. Style, as a separate trait, contributed most to the final judgement of the text, with 4%. Also *introduction* and *lay-out* explained some variance, both with 3%. *Transformation to an article, title and imitation* did not explain any variance in this analysis.

Of the combination traits, *audience* contributed most to the total explained variance with 4%. Also in this analysis the communal contribution of the higher order concerns was quite high, but compared to the expert raters it was somewhat lower: 14 percent. Even though this number was not as high as with the expert raters, it is still a substantial contribution. Also for the grade 7 students it was hard to distinguish between the higher order concerns and the traits overlapped when explaining the unique variance.

*Layout* and *spelling and punctuation* explained more variance than the variance in the expert ratings. *Layout* uniquely explained 3% of the variance and *spelling and punctuation* contributed 2% to the final judgement. The communal effect of the lower order concerns when *lay-out* and *spelling and punctuation* have been subtracted was 2 percent.

*Common* in table 9 is what cannot be attributed to the separate traits or combinations. A very high contribution of *common* remains: 0.21. Apparently the traits, including the lower and higher order concerns, were overlapping. The grade 7 students apparently had difficulties distinguishing between the traits.

These regression analyses showed that the communal contribution of the higher order concerns, style and audience, contributed most to the final judgement of the text. *Layout* and *spelling errors and punctuation* predicted the final judgement more than when expert raters rated the texts.

### 4.5.2.3 Comparison Trait Contribution from Grade 7 Student Raters and Expert Raters

When the traits from grade 7 student ratings and expert ratings were compared for their contribution to the final judgement, the traits in the expert raters explained more variance (77%), than the traits of the grade 7 student raters (59%). Which means that 23% of the variance that contributed to the final judgement remained unexplained in the assessment model rated by the expert raters. While 41% remained unexplained for the ratings by the grade 7 students. This could suggest that the expert raters rated more consistently and rated more according to the assessment model than the grade 7 students.

The traits *style, introduction* and *lay-out* in the assessment models of grade 7 student raters contributed more to the final judgement than these aspects did when the texts were rated by the expert
raters. The language of the text (*audience language*) was also of more importance in the traits used by the grade 7 students than in the traits in the assessment model of the expert raters. The traits *transformation to an article* and *end* in the assessment models by the expert raters explained more variance than they did in the traits filled in by the grade 7 students.

Other differences were that the grade 7 students appeared to have difficulties distinguishing between the lower order concerns and higher concerns, judging by the high contribution of *common*. The expert raters had merely a small contribution of *common*.

### 4.6 IMPROVEMENT FROM CONCEPT TO FINAL TEXT

The freshmen students wrote a concept text for the children’s page and received feedback from their peers (2 reports) and from the authentic audience (2 reports). And using this feedback they wrote the final text of the children’s page. Subsequently, expert raters and grade 7 students rated the concept text and the final text. To evaluate the texts, they used an assessment model. This assessment model included traits in which they had to indicate their opinion on a 5-point Likert scale and they had to provide an overall grade for the texts on a scale of 1 to 10.

The concept version was assessed by two grade 7 student raters and two expert raters. The final version of the text was also rated by four raters; these raters were different raters, thus, the concept text and the final text have not been assessed by the same raters. Also in this analysis the ratings of the grade 7 students and the expert raters were regarded separately, as they rated more consistently separately than combined.

The difference between the concept text and the final text and if this disparity was significant was analyzed using a paired-samples t-test. The first test was conducted using the ratings rated by the expert raters. This test indicated that there was no significant difference in text quality when concept version and the final version of the text were compared ($t(19) = 0.42; p = 0.68$). The rating of the concept texts, however, was slightly higher ($M=6.5$, $SD=1.24$) than the score on the final version of the text ($M=6.4$, $SD=1.45$). This analysis showed that, according to the expert raters, the students did not significantly improve their texts, when the concept version was compared with the final version of the text.

Also the ratings by the grade 7 students were analyzed. As the data was not normally distributed, no paired-samples t-test could be conducted. Therefore a non-parametric test, namely a *Wilcoxon Signed Rank Test* had been conducted. This test revealed that there is no significant difference between the ratings on the concept texts and final texts: $z = -1.83$, $p < 0.07$. The median score did increase from the rating on the concept text ($Md = 6.5$) to the rating on the final text ($Md = 7$). Also this analysis seems to indicate that there was no significant improvement from the concept text to the final text. One would expect that after receiving four feedback reports the writers would improve their texts when revising. To get a better picture of the ratings of the text and these results the judgements of all raters on all 20 texts is represented in table 10. Table 10 also seems to suggest that the expert raters and the grade 7 students did not agree on the quality of the texts. Quite often the expert raters claimed that a text had improved in quality, while the grade 7 students thought it had deteriorated, or vice versa. This could indicate that the expert raters and grade 7 students have a different view of what a text of good quality is.
Even though there did not seem to be a communal improvement from concept text to final text, either for the raters within or between the rating groups, the judgements seem to indicate that there were considerable changes in quality from concept text to final text per text. These disparities were in my view due to the individual lenience or strictness of the raters, as the inter-rater agreement between the raters (both the expert raters and the grade 7 raters) was usually of moderate agreement (see section 4.4). Thus, there were probably hardly any quality differences between the concept texts and the final texts. The differences were most likely caused by the quite inconsistent ratings of the raters. These ratings were presumably a side effect of the assessment method. As the concept text and the final text of one writer have not been assessed by one rater, but by various raters. This warranted the reliability of the evaluation, as no comparison can be drawn. However, this method had as a drawback that small quality differences between versions could not be discerned. In other words, the judgements from concept to final text were somewhat inaccurate. Another argument for the view that the texts plausibly had not changed in quality was the usage of revisions. The freshmen students did not revise much; on average they made 16.7 revisions. And when they made changes in their concept texts, the revisions were mostly local changes and simple operations like additions or substitutions. These changes were most probably not responsible for quality differences. This view is further enhanced by the more thorough discussion of the revisions in section 4.8.

Table 10: Difference in judgement of the improvement of the texts (from concept to final text) by the raters (two expert raters and two grade 7 student raters per text).

<table>
<thead>
<tr>
<th>Text</th>
<th>Expert raters (2 raters per text)</th>
<th>Expert raters (2 raters per text)</th>
<th>Grade 7 students (2 raters per text)</th>
<th>Grade 7 students (2 raters per text)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average rating</td>
<td>Difference concept and final</td>
<td>Average rating</td>
<td>Difference concept and final</td>
</tr>
<tr>
<td></td>
<td>Concept → Final text</td>
<td></td>
<td>Concept → Final text</td>
<td></td>
</tr>
<tr>
<td>Text 1</td>
<td>7.5 → 9.0</td>
<td>+1.5</td>
<td>6.0 → 8.0</td>
<td>+2</td>
</tr>
<tr>
<td>Text 2</td>
<td>7.0 → 5.5</td>
<td>-1.5</td>
<td>5.0 → 7.0</td>
<td>+2</td>
</tr>
<tr>
<td>Text 3</td>
<td>8.0 → 7.5</td>
<td>-0.5</td>
<td>8.0 → 8.0</td>
<td>0</td>
</tr>
<tr>
<td>Text 4</td>
<td>4.5 → 8.5</td>
<td>+3.5</td>
<td>5.0 → 6.5</td>
<td>+1.5</td>
</tr>
<tr>
<td>Text 5</td>
<td>5.0 → 6.0</td>
<td>+1</td>
<td>5.5 → 6.0</td>
<td>+0.5</td>
</tr>
<tr>
<td>Text 6</td>
<td>4.5 → 4.0</td>
<td>-0.5</td>
<td>6.0 → 6.5</td>
<td>+0.5</td>
</tr>
<tr>
<td>Text 7</td>
<td>7.0 → 7.0</td>
<td>0</td>
<td>8.0 → 6.5</td>
<td>-1.5</td>
</tr>
<tr>
<td>Text 8</td>
<td>5.0 → 6.0</td>
<td>+1</td>
<td>6.0 → 6.5</td>
<td>+0.5</td>
</tr>
<tr>
<td>Text 9</td>
<td>7.0 → 7.0</td>
<td>0</td>
<td>6.5 → 7.5</td>
<td>+1</td>
</tr>
<tr>
<td>Text 10</td>
<td>7.0 → 7.5</td>
<td>+0.5</td>
<td>7.0 → 7.0</td>
<td>0</td>
</tr>
<tr>
<td>Text 11</td>
<td>6.5 → 7.0</td>
<td>+0.5</td>
<td>8.5 → 7.0</td>
<td>-1.5</td>
</tr>
<tr>
<td>Text 12</td>
<td>5.0 → 6.0</td>
<td>+1</td>
<td>8.5 → 6.5</td>
<td>-2</td>
</tr>
</tbody>
</table>
4.7 CONTRIBUTION OF REVISION AND FEEDBACK ON TEXT QUALITY

In the literature review it was explained that various factors can contribute to text quality, for instance the beliefs a writer has about writing, but also revision can contribute to writing quality. In this section the effect of various variables on text quality will be described. Firstly, the contribution of the quality of the concept version on the final version will be elucidated; secondly, the contribution of revising more on text quality is explained. Then the contribution of receiving more feedback on the number of revisions will be analyzed. And last, the contribution of implementing more feedback, implementing more feedback from peers or from grade 7 students will be described.

To study the influence of these variables simple regression analyses were conducted. This analysis method investigated - using a correlation between the independent and dependent variable - if there was a predicting connection between those variables.

4.7.1 CONCEPT VERSION

The quality of the concept text can predict the quality of the final text. Also in this analysis the ratings by the expert raters and the grade 7 raters was conducted separately. For both of the rating groups the B-values were not significant. This means that the rating on the concept version did not predict the rating on the final version of the text, even though one would expect it to predict the final score on the text.

4.7.2 REVISION

Former studies on the influence of revision on text quality seem to agree that revising can decrease the quality of the text (Van Gelderen, 1997, Stevenson, Schoonen and De Glopper, 2006). In this analysis the contribution of revising more on text quality was analyzed. For both the expert rater ratings and the grade 7 rater ratings, there was no correlation between the number of revisions and the final judgement.
of the text. The B-values were not significant in the regression analysis, which means that the number of revisions did not make a contribution to predicting the final rating of the text.

### 4.7.3 FEEDBACK AND REVISIONS

It could be assumed that when students receive much feedback on their text; they will also revise more. The following regression analysis investigated this assumption. The analysis showed that neither of the B-values was significant, this means that the predictor (number of revisions) did not contribute to predicting the number of revisions.

### 4.7.4 IMPLEMENTING FEEDBACK

The relationship between the implementation of the feedback and text quality is not straightforward. Feedback itself cannot contribute to the quality of the text: revision is the mediating factor and will most probably contribute more to the quality of the text. The relationship between feedback, revision and text quality is therefore difficult. Revision is also a confusing variable, as it is uncertain how it contributes to text quality. Even though more experienced writers revise more, revising more not necessarily means that the text will be of better quality (Stevenson, Schoonen and De Glopper, 2007). Students, who have written a bad concept text, will receive more suggestions for improvement than students who wrote an excellent first text. The students who wrote a poor text will therefore probably revise more, but not necessarily improve their texts, or receive better ratings than the students who wrote a good text in the first try. The contribution of feedback and revision on text quality is therefore quite hard to analyze.

Even though there is no straight relationship between feedback and text quality it is still important to examine the contribution of the number of implemented feedback on text quality. It is significant, however, to bear in mind that revision is mediating between these two variables. To analyze whether the number of feedback that is implemented (feedback from both grade 7 students and expert raters) can contribute to the final rating of the text, two regression analyses have been conducted. For the ratings of both the grade 7 students and the expert raters the amount of implemented feedback did not contribute to the final rating of the text.

In the former thesis the reasons for implementing the feedback from the grade 7 students and the expert raters was analyzed. One part of this thesis was to investigate whether students who had implemented more feedback from their peers or the grade 7 students, also wrote better texts of better quality. The students declared in the reflection reports that they perceived the feedback from their peers as more helpful. Two regression analyses, with the ratings from the grade 7 students and the expert raters separately, revealed that if writers implemented on average more peer feedback, this could not predict the final rating of the final text.

One could argue that if the implementation of peer feedback does not increase text quality in between drafts, taking up the feedback from the grade 7 students could improve the text, as these
students are the authentic audience of the text. A simple regression analysis showed that implementing more feedback from the grade 7 students using the ratings from the grade 7 students did not predict the final rating of the text. However, when only the ratings from the expert raters were used; it seems as if implementing more feedback from grade 7 students contributed to the quality of the text. This regression analysis showed that both B-values were significant, this means that they were different from 0. Thus, the average implementation of G7 feedback could predict the judgement of the end version of the article. Table 11 below shows that 37 percent of the end judgement variance can be attributed to the average implementation of G7 feedback comments \((Adjusted \ R^2 = 0.37)\).

<table>
<thead>
<tr>
<th>Table 11: Prediction of the final judgement rated by the expert raters of the text with the predictor “implementation of feedback comments from the grade 7 students”</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B</strong></td>
</tr>
<tr>
<td>Constant</td>
</tr>
<tr>
<td>Average of implementing Grade 7 feedback</td>
</tr>
</tbody>
</table>

Note \(Adjusted \ R^2 = 0.37 \quad *p<.05\)

Table 11 revealed that when students did not implement feedback from the grade 7 students \((X = 0)\), the model predicted that they will obtain a rating of 5.25. The other B-value, 0.03, indicated that if the predictor variable was increased by 1, this will yield an increase of 0.03. Thus, when students would on average implement 1 feedback comment of the grade 7 students, the model predicts that the score on the end version of their text will increase with 0.03 points. In other words, if a student received feedback from the grade 7 student and did not use any of the comments, he would receive a 5.25 as a final score according to this model, however, if he would implement 1 feedback comment more, he would receive a 5.28, as his grade would increase with 0.03 points.

4.8 REVISION STRATEGIES FRESHMEN STUDENTS

As already alluded to in section 4.6, the freshmen students did not revise much. The minimum number of revisions executed by the students was namely 1 revision and 50 revisions was the maximum. The mean was 16.7 and the median was 9. These were all revisions, including formal changes, i.e. correcting spelling errors. These small numbers seem to indicate that the students did not revise thoroughly, but changed only what was really necessary.

Also the scope of the revisions was examined; thus whether the writer had focused on for instance the word level or the sentence level when employing revisions. The revisions could have five different types of scope: surface (indicated with the number 1); word (number 2); sentence (number 3); paragraph (number 4) and text as a whole (number 5). The minimum scope was 1 (surface level) and maximum was 4 (paragraph level). This meant that none of the writers have revised the text as a whole.
Thus, the writers did not make changes in their texts that altered their texts as a whole. The mean of the revisions the freshmen students have made was 2.3, which was somewhat above the word level. The median of the scope of the revisions was 2, which was word level. These figures indicated that the writers mainly have revised on the word level.

It was also examined whether the students have made meaning altering changes to their texts. There were four categories. The first category was formal changes, which for instance implied that the writer changed the spelling of a word that could be spelled formally and informally. The second category was meaning preserving for micro changes, the writer for instance used a synonym that did not change the meaning of the concept text. The third category was: meaning altering affects the concept text, a revision made by a student changed the quality of the concept text and the fourth category was meaning altering affects the summary, by changing this aspect it changed the summary of the text. The changes the students made mostly did not change the meaning of the text, which were the categories form and meaning preserving for micro changes combined: these two categories included 293 of the 334 changes, which was 87.8 percent of all the changes made in the texts. The remaining 12.3 percent of the changes, changed the quality of the concept text, but it did not change the summary of the text. This means that no student has made a very thorough revision that changed the summary of the text.

As was described in the literature review of this thesis, the types of revisions that were distinguished were based on the typology of Faigley and Witte (1981). These categories were: addition, deletion, substitution, permutation, distribution and consolidation. The freshmen students substituted most frequently (a proportion of 0.29 on average), they also added a lot of new information (a proportion of 0.27 on average). They also deleted words, sentences or paragraphs in their texts (a proportion of 0.14 on average). They hardly shifted parts of the sentences in a different manner (permutation): a proportion of 0.06 on average, they also did not put a sentence into two sentences (distribution): a proportion of 0.03 on average, nor did they put the information that was included into two sentences into one sentence (consolidation): a proportion of 0.004 on average.

Also the improvement, deterioration or equality of every single revision has been studied. Thus, the quality of the final text as compared to the concept text has been studied per revision. In total 334 revisions have been made by the 20 freshmen students. Of these revisions 190 changes were improvements (56.9 percent), 22 changes were deteriorations (6.6 percent) and 122 changes did not change the quality of the text (36.5 percent). Thus, most of the students´ revisions were improvements, but also more than a third of the revisions did not change the quality of the texts.

4.9 ANECDOTAL MATERIAL OF WRITERS REVISINNG

In the last section it was explained that writers hardly revised their texts for this assignment. There were a few students who did revise their text. In this section some anecdotal material will be provided of a student who revised successfully and a student who deteriorated her text by revising.

In table 10 it could be seen that text 4 has improved conspicuously. The text improved with 3.5 points for the expert raters (from 4.5 to 8.5) and with 1.5 points for the grade 7 student raters (from 5 to 6.5). When examining the assessment model filled out by the four raters, it mostly received negative comments: they commented that the text was not suited for this audience, as it was too childlike; the text also lacked humor, it contained too many grammatical errors, the text was pedantic and neither the
title nor the introduction was attractive. The feedback providers who provided feedback on the concept text for the benefit of the writer, also had many comments. They commented on similar aspects of the text: on the childlike wording, the grammatical errors and the unattractive title. The writer took this feedback into account and made 21 revisions, of which 14 revisions were not based on the feedback, but were self-initiated. She changed, among other things, the title, ameliorated the grammatical errors and deleted or substituted the childlike phrases. The amelioration of the grammatical errors can increase the appreciation of the text, as Jansen (2010) pointed out: if a writer consistently makes the same grammatical errors, readers perceive the writer as incompetent and they do not trust the content of the text. The alternation of the title also had a positive influence on the text, as this title was tailored more to the audience of 12/13 year old children. The title of the concept text was: “Writing is fun”, which was changed in the final text into: “Copying is not that bad!” The writer also seemed to consider her audience as she deleted or substituted phrases as: “Difficult word huh, observe?” and “As also Miss wants all children to write well”. These phrases probably negatively affected the perception of the feedback, as it was presumably too childlike and therefore underestimated the readers. By deleting these sentences, the author took the audience more seriously; which might have had a positive influence on the quality of the text. The revisions the writer made are mainly focusing on the audience and resulted in better ratings on the final text. This could suggest that when raters rate a text, they chiefly concentrate on audience orientation.

Another text that got a neutral or negative judgements by the expert raters and the grade 7 students was text 7. The expert raters kept the text of similar quality and provided the text with a 7 for both versions. The grade 7 students, however, provided the concept text with an 8 and the final text with a 6.5. Text 7 was about the differences between the speeches the Dutch queen Beatrix gave in Indonesia and Israel. The structure of this text was understandable, even though sometimes the transitions from the speech in Indonesia to the speech in Israel seemed not to be marked clearly enough. This writer also seemed to underestimate her readers, providing unnecessary information that children of that age probably already knew: “those countries are very far away, so she had to go there by plane” and “[…], that is called a state visit. She wears very beautiful clothes and she usually wears a beautiful, expensive hat.” Despite these demerits, it seemed to me that this concept text is of above moderate quality.

This student did not receive very severe feedback: her peers said that her text was somewhat childlike; the grade 7 student feedback providers did not like the topic, they also thought it was too childlike and they missed humor. The children seem critical, but these students almost always said that they did not enjoy the theme of the text. This writer took note of the feedback that her style was too childlike and changed this in her final version. She deleted or substituted 9 childlike phrases. There were 43 revisions in total, 34 of those were not based on feedback but were self-initiated. In these revisions, the writer added, substituted, deleted and rearranged. She also made a few formal changes. She mostly changed words and sentences, but there were also some paragraph changes. This writer had added two new paragraphs at the commencement of the text; paragraphs that both function as the introduction. In the first paragraph she explained the relation between the Netherlands and Indonesia, in paragraph two the connection between Israel and the Netherlands was elucidated. She probably wanted to separate both countries to make the differences between them clear. However, an introduction that has two paragraphs would appear to be unconventional and confusing. To create those two new paragraphs she added some new sentences, but she also used sentences from other paragraphs of her concept text. Paragraph 6 has also been transformed. Initially this paragraph described the emotional charged connection between the Netherlands and Indonesia. This information is moved to paragraph one in the
final text. Because the writer had already explained the connection between these countries at the start of the text, she probably thought she was able to immediately describe the consequences of the weak relationship between the Netherlands and Indonesia on the characteristics of the speech queen Beatrix gave in this country. This however caused processing problems, as the information of paragraph one is too far away. As the writer had rearranged the information of the concept text, the structure was incoherent: the text, in my view, was made up of isolated sentences and lacked a logical exposition of the topic. This was probably a direct result of the self-initiated revisions by the writer.

When I discussed the merits and demerits of the concept text of this student, I opinionated that the concept text did not distinguish clearly enough between the speeches in Indonesia and Israel. Apparently, the student of this text also discerned that the distinction between Indonesia and Israel could be marked more clearly and she decided to ameliorate her text on this aspect. As discussed in the introduction, it is strenuous to revise your text. According to Sadler (1989) there are three conditions writers have to internalize before they are able to revise successfully. Condition one of his model is that the writer knows what a good text is; the desired norm. The writer of text 7 apparently knew what the desired norm was, as she wanted to improve her text. She was also capable of comparing her own text to this desired norm, and concluded that her text could be improved in terms of structure. Being able to evaluate your own text as compared to a good text is the second condition of the model of Sadler. Her revisions however did not improve the structure of the text, but deteriorated it. This means that the writer had not internalized condition three, that implies that the writer knows which steps need to be taken to close the gap between the current level and the desired norm. Revising the structure of the text, thus, revising globally, proved to be too complicated for this student. Sommers (1980) also claims that writers do not have the proper strategies to revise their texts globally. Her revisions did not improve the text, on the contrary, it had severely damaged the ‘flow’ of the text and has deteriorated in quality considerably.
5. DISCUSSION

5.1 INTRODUCTION

In this chapter a summary of the results will be given in section 5.2. In section 5.3 the results will be connected to former research and the differences and similarities will be discussed. Subsequently, the method of this research will be discussed in section 5.4. In 5.5 the relevance of this study for society will be elucidated and finally, it is described in 5.6 which questions are raised by this study and what should be studied in the future.

5.2 SUMMARY RESULTS

In this section answers will be provided for the five research questions. They will all be answered separately.

1) What is, according to the research literature, the most valid and reliable method to assess popular science articles for a young audience?

In this thesis I had to choose the ‘best’ assessment method for this particular assignment and for both types of raters. Thus the assessment method had to be suitable for expert raters and grade 7 raters. After close examination of all the literature it seemed that using comparative methods would have been the most reliable option. If this would have been used in combination with an analytical and multiple trait model, this would have yielded the best results. However, this option was not chosen, as this method would be too strenuous for the grade 7 raters, and would probably be also somewhat too difficult. Comparative methods also would have taken too much time of the raters. Thus, reliability was somewhat sacrificed, favoring practical issues. I chose to use a combination of an analytical and multiple trait model, as this provided diagnostic information about the strengths and weaknesses of the texts. Another reason to use an analytical scale is that Shohamy, Gordon and Kraemer (1992) found that inexperienced raters can rate reliably when using an analytical scale.

The combination of an analytical and multiple trait model was sufficient for this particular assignment with the time and raters that were available. The aim of this study, to find the model to assess texts for popular science articles and a young audience, appeared to be too ambitious. An assessment model usually changes when another assignment is given, thus to develop a model that can be applied to every assignment focusing on popular science articles written for a young audience, was not feasible.

However, some information can be given about what the most important features are when assessing texts for children. After careful consideration of the texts and the ratings, and hearing the opinions of the raters about suitable texts for children 12/13 year of age, it seems as if not underestimating the audience is very important for an assessment model for a text written for children. All the raters, when asked for their perception, pointed this out and also in the ratings it became clear that underestimating your writer decreases the appropriateness of the text for this audience. In an
assess the texts than the expert raters. However, caution should be exercised in drawing these kinds of conclusions from these regression analyses: the focal points of these raters are not directly studied. It is

2) Which traits contribute most to the final judgement of the text, as rated by the expert raters and the grade 7 students?

For the expert raters 77% of the variance was explained by the assessment model, this means that 23% of the variance remained unexplained. The communal contribution of the higher order concerns (when all the separate traits, for instance structure and the two combined traits – audience and imitation - are subtracted) contributed most to the final judgement of the text, with 57%. This might mean that the expert raters had difficulties distinguishing between the separate higher order traits, and thus that the traits overlapped. Of the separate traits, the end attributed most to the final judgement of the text with 4%. Structure, introduction and spellings errors and punctuation did not uniquely explain any variance. When all the separate traits, the combined traits, the higher order concerns and the lower order concerns are subtracted from the total explained variance, the common contribution remains; this is a unique and common contribution. Common is what cannot be attributed to the separate traits or the combinations. The contribution of common is quite small for the expert raters: not a lot of variance remains after all other separate traits and communal contributions are subtracted.

For the grade 7 raters 59% of the variance can be explained by the assessment model. 41% of the variance remains unexplained. This suggests that the ratings of the expert raters are more based on the assessment model than the ratings of the grade 7 students. The communal contribution of the higher order concerns is 14%, this is somewhat lower than for the expert raters, but it is still a substantial number. This means that also the grade 7 raters had difficulties distinguishing between the separate traits and that there still remained a high contribution of communal variance. The separate traits rated by the grade 7 students that contributed most to the explained variance is style, introduction, lay-out and the communal contribution of audience. Spelling errors and punctuation also contributed to the total explained variance, while this trait did not contribute to the final judgement for the expert raters.

In sum, the traits in the assessment model when rated by the grade 7 raters explain less variance than when the traits are assessed by the expert raters. This could indicate that the ratings by the grade 7 raters are more based on aspects outside the assessment model than for the expert raters. The commonality analysis also suggests that the grade 7 students have difficulties distinguishing between the traits and that the traits have overlapped. For both the expert raters and the grade 7 raters the communal trait audience contributed to the total explained variance. There are also differences. Introduction, lay-out and spelling errors and punctuation contributed more to the total explained variance when rated by the grade 7 students, than when rated by the expert raters. For the expert raters the most explained variance is in the communal contribution of the higher order concerns; this means that the separate traits overlap and the raters have difficulties distinguishing between the traits of the higher order concerns.

These contributions from the traits to the final judgement can also be seen as the focal points of the raters. Expert raters then focused most on the common higher order concerns and the grade 7 students concentrated more on introduction, lay-out and spelling errors and punctuation when rating the texts than the expert raters. However, caution should be exercised in drawing these kinds of conclusions from these regression analyses: the focal points of these raters are not directly studied. It is
not known what these raters have actually based their judgements on. These regression analyses, however, do provide an impression of what the focal points of the grade 7 raters and the expert raters are. But to actually be able to study the focal points of raters, think-aloud protocols, or a similar technique, should be used.

3) What is the relationship between the final judgements of the texts and the (implementation of the) feedback received from two types of feedback providers?

The number of received feedback comments does not contribute to text quality. Implementing more feedback from the expert raters also does not contribute to the final judgement of the text. However, when raters implement more feedback from the grade 7 students, this can predict the final rating of the text. When a writer implements 1 feedback comment more from grade 7 students, the writer will increase his score on the text with 0.03 points. This result could suggest that receiving feedback from grade 7 students on an assignment focusing on this audience might be more beneficial for text quality than receiving feedback from peers. It should again be pointed out here that the relationship between feedback and text quality is not at all straightforward: revision is mediating between these variables. Also other variables can influence text quality: writing apprehension, self-efficacy, writing beliefs and writing approach, among other things.

4) What are the features of the revisions made by the freshmen students and what is the relationship between the number of revisions and the final judgements of the texts?

The freshmen students hardly revised their texts, as on average they only made 16.7 changes. The scope of these changes was also quite small: somewhat above word level. The students apparently mostly revised locally. Also the type of operations that was executed by the freshmen students was quite simple; they mostly substituted (for instance using a synonym), added new information or they deleted information. They hardly ever used the more complicated operations permutation, consolidation and distribution. It was also analyzed whether the writers changed the meaning of their text. They hardly did: 87.8 percent of the changes did not change the meaning of the text and therefore were minor revisions. The remaining 12.3 percent changed the meaning of the concept text, but no writer revised extremely intensively, as no writer changed the summary of the text. This analysis shows that the writers hardly revise their texts, even though most of these students have received many feedback comments from their peers and the grade 7 students. If the students revise, they use simple operations and revise locally.

The anecdotal material of two revising students shows that revising texts is difficult. One writer is able to revise her text successfully, even though she mostly revises locally. The other student who found a problem in the structure of her text, made more rigorous changes to her text. She was, however, unable to successfully solve her organizational problem. Her revisions did not improve her text, but deteriorated it.

Also the relationship between the number of revisions and the final judgements of the texts has been studied. A regression analysis showed that the number of revisions do not contribute to the final judgement of the text. This outcome is not surprising, as the students hardly revised their texts, and if they did, they revised locally, these types of revisions will probably not alter the quality of the texts.
During the development of the assessment model, reading the texts, ratings and feedback forms and hearing the perceptions of the raters what a suitable text is for 12/13 year old children, suggested that not underestimating your audience is the most important feature of a text for a young audience. This information cannot be linked to former research, as the studies on assessing writing do not provide any information about how to develop a reliable and valid assessment method to assess popular science articles for a young audience. The research on assessments mostly focuses on developing the most reliable and valid method in general, regardless of focusing on audience or genre.

The contribution of the traits showed that there were big differences between the expert ratings and the grade 7 ratings in the way the traits contributed to the final judgement of the text. Cumming (1990) found that novice raters were mostly editing as compared to the expert raters. Also in this study the trait spelling errors and punctuation in the ratings by the grade 7 students contributed more to the final judgement than by the ratings of the expert raters. The surface aspect lay-out also contributed more to the final judgements for the grade 7 students than for the expert raters. I could not find any study that investigated the influence of this text feature on final text quality. De Glopper (1988) did find that teachers also sometimes focus on superficial features when assessing texts. In his study some teachers were leaded in their judgements by spelling errors and punctuation, while others focused on text length. His study suggests that different raters also concentrate on different aspects. This claim is also made by Lumley (2002) who states that if an analytical scale is used, all raters will differ in their focus on the various traits. These two studies seem to suggest that rater variability is substantial.

The grade 7 raters also had difficulties distinguishing between the traits, judging by the high number on common. The traits for them overlapped. The total explained variance was also considerably lower for the grade 7 students: 59% as opposed to 77% for the expert raters. This seems to suggest that the grade 7 raters focused on features outside the assessment model. This claim could correspond to the study by Cumming (1990) who found that novice raters were basing their judgements on intuition.

The contribution of feedback and revision to the final judgement of the text hardly yielded any results. This does not match up with former studies on revision. Revision usually affects text quality (Breetvelt, Van den Bergh and Rijlaarsdam, 1994; Van Gelderen, 1997; Stevenson, Schoonen and De Glopper, 2006). Receiving more feedback did not contribute to text quality and implementing more feedback from peers also did not attribute to the final rating of the text. However, implementing more feedback from grade 7 students did contribute to the final judgement. This is in line with former research, as interacting with the audience can improve text quality (Sato and Matsushima, 2006).

The analyses of the revisions by the freshmen students showed an image of students who hardly revise, and if they do they revise locally. This result is in line with former research. Faigley and Witte (1981) also found that the inexperienced writers, thus non-professional writers like students, mostly focus on local changes. Also Van den Bergh and Meuffels (2000) claim that more competent writers make changes to bigger fragments of the text, than less competent writers. Also Dave and Russell (2010), Fitzgerald (1987), Lai (1986) and Sommers (1980) found that less experienced writers revise locally.

Faigley and Witte (1981) found that inexperienced writers revise less than more experienced writers. This is in line with the results of the analyses in this thesis, as the students did not change much in their texts.
Another finding in this study is that students hardly changed the meaning of their texts (12.3% of all revisions), this matches up with Faigley and Witte (1981). They found that inexperienced writers changed the meaning of their text in 12% of all revisions.

The operations the freshmen students use in this study when revising are simple; they mainly used substitutions, additions and deletions. Lai (1986) also found that students mainly used substitutions, additions and deletions. This is also in line with former research by Faigley and Witte (1981) who claim that inexperienced writers use these operations, while expert writers mainly applied additions, consolidations and distributions.

The anecdotal information about students revising indicates that revising the structure of the text was particularly hard for one student. This is also in line with former research. Sommer (1980) namely claims that students lack proper strategies; they know how to revise locally, but they do not know how to revise a text as a whole.

All these revision results taken together seem to indicate that students do not want to take the time or effort to revise their texts thoroughly, as they only make simple revisions. This seems contrary-intuitive, as they are incited to revise their texts, by providing the students with feedback and also by asking them to hand in a concept text and a final text. They receive all the means to revise their texts. But they hardly revise. Not willing to take time and effort to intensively revise the text might be part of the reason. However, the reason might also be that students are not yet capable of making intensive revisions: they might not have the proper strategies to actually revise their texts thoroughly (Sommers, 1980). Sommer’s view is supported by the anecdotal material: the one student who made very intensive, global revisions deteriorated her text. She was not able to fill the gap between her performance and good performance.

5.4 METHODOLOGICAL DISCUSSION

In this section a few methodological decisions will be discussed and which effect they had on the results of this study.

The reliability and the validity of the assessment method were discussed partly in the results chapter. However, also in this methodological discussion the assessment procedure will be discussed extensively, as some things could have been improved. Most improvements that will be put forward here are methodological issues that could not have been changed, due to time and resource restrictions. A balance between reliability, validity and practicality was found for this assessment procedure. Most decisions that have been made for this assessment procedure, I would make again, regardless of the disadvantages. The methodological issues raised here, therefore mainly serve as recommendations for future assessment procedures.

In this study an analytical/multiple trait model was chosen: it provides diagnostic information and it is easy to use for the young grade 7 raters. However, by choosing (only) this method the reliability is somewhat sacrificed, as also using comparative methods would have made the assessment method more reliable. In this study it was not possible to also use comparative methods for practical reasons. The grade 7 raters and the expert raters did not have that much time to rate and using comparative methods would probably be too strenuous and difficult for these young raters.
The results showed that the inter-rater agreement between the raters was not that high. This was probably due to the lack of proper rater training. The raters have been trained to rate the texts, however, as they all had little experience rating texts, the raters should have been trained somewhat more thoroughly. Preferably by rating some texts together, then letting them rate a few texts individually and then discussing the differences. However, as the training of the expert raters and the grade 7 raters had to be equal, an extensive training was not feasible, as the grade 7 raters rated the text in class and there was unfortunately no time for a thorough rater training. The assessment literature also points out that practicality issues are inevitable in assessment procedures. Weigle (2002, p. 48) points out that it is “virtually impossible” to satisfy all criteria in developing an assessment.

Another problem with the inter-rater agreement was that this agreement could not be analyzed well, as the raters did not share enough common texts. This problem was also due to practicality: the expert raters were very busy, thus they did not have the time to assess many texts. And as the teacher did not have that much time for this research, the grade 7 students only had one class to be trained and to rate the texts. This means that it was not possible for these raters to rate more than two texts.

Another decision in the rating process that had big consequences is the choice of having the concept text and the final text by the same writer be rated by different raters. The idea is that the raters would provide an unbiased opinion of the text. However, as they were unbiased, they also did not find small improvements or deteriorations from concept text to final text. Thus, the improvement (or deterioration) of the text cannot be examined properly. It would have been a better choice in this study to let the raters rate both the concept text and the final text.

The assessment model itself seemed to be sufficient for this purpose. The traits were specific enough and using five bands was the correct choice as the raters were able to distinguish between them. However, the trait about structure was quite confusing for the raters, as the text had to be logical, but also had to have the atypical structure of Menno Steketee’s style. It would have been better to have explained this trait even clearer. A trait that should have been added on the assessment form should have been: “On a scale of 1 to 10 how would you grade the text by Menno Steketee?” This trait should have been added, as the grade 7 students did not really appreciate the style by Menno Steketee. Thus, their opinions could have affected the results on the ratings.

The number of feedback implementations was analyzed using the opinion of the writers in their reflection reports. It would have been better to have compared the feedback comments to the feedback implementations, as it is uncertain if their information corresponds to the actual implementation. However, this analysis was not feasible due to the time restrictions of this thesis.

The revisions are categorized by only one researcher, which decreases the reliability of this method. However, if in doubt, a second researcher was consulted. For future research it would be best to have two (or more) raters categorizing the revisions. As revision was not the aim study of this thesis, there was no time, nor resources to have more raters categorizing the revisions.

The revisions of the raters were studied by looking at their writing products: feedback reports, concept texts, final texts and the ratings. To actually study a writer’s process, an ‘online’ method could have better been used, for instance by using think aloud protocols and/or key stroke logging. This is, however, not considered, as the aim of this study was not to study revision processes, but mostly to develop a reliable and valid assessment model. Thus, there was no time to investigate the revision process.

In this thesis a commonality analysis has been conducted. I would like to point out that as this study is not experimental in nature, and no variable is manipulated, no causal conclusions can be drawn. Every rating depends on the text features that are ‘coincidentally’ present in the accidental samples. To
draw causal relations the text features have to be manipulated. However, the contributions of the traits do provide an impression of the focal points of the raters.

5.5 SOCIETAL RELEVANCE

The variation in the contributions of the traits to the final judgement can be of importance to people who assess writing. It provides some insight into the focal points of two types of raters and also contributes to the view that a lot of variance still remains unexplained. This information might help developing new rater trainings, to help them focus more on the assessment model.

The information that receiving feedback from the authentic audience can contribute to text quality, might be of importance to writing teachers who want their students to improve their writing ability. Also the features that are important when assessing a text written for 12-13 year old children, could be of interest for language teachers. They might want their students to write a text for an authentic audience and they might need guidelines how to assess these types of texts.

The findings regarding revision shed some more light on the difficulties students have with revising globally. This result might indicate that language teachers should aid their students with acquiring the proper skills to revise globally, as this study suggests that they are not able to now.

5.6 FUTURE RESEARCH

This thesis provides some insight into the influence of text features on text quality. As this study was not experimental, the text features were not manipulated: the aspects were merely accidentally present in the samples. However, to be able to draw causal conclusions between text features and text quality, text aspects need to be manipulated. This would shed light on the relationship between text aspects and text quality.

In this study the focal points of raters are only studied indirectly using regression analyses. It would also be very interesting to study focal points more directly and thoroughly, for instance by using think aloud protocols. This would shed light on the text aspects raters base their final judgement on. This would be of importance for future assessment research, as it provides information on how raters should be trained to focus on certain aspects; to be able to increase the reliability of assessment.

It would also be interesting to more specifically focus on these young raters and their ability to rate texts. For instance by setting up a longitudinal study and observing their development in rating texts.

In this study it was found that implementing more feedback from the authentic audience can predict the final judgement of the text. This result suggests that there is a positive causal relation between using more feedback from the authentic audience and text quality. However, in future research the influence of receiving feedback from a young audience on text quality should also be examined experimentally. All participants should then write for a young audience, but not every group receives authentic feedback. This would actually provide insight in the effects of receiving feedback from
an authentic audience. The reflection reports written by the students, however, already give some indication that receiving feedback from an authentic audience is useful: “The feedback from the authentic audience is more useful than the feedback from the peers, as only then you know if your text is successful and if they like to read it.”


In de rubriek Natuuuuuurlijk!!! op de kinderpagina van NRC Handelsblad schreef Menno Steketee de afgelopen jaren op gezette tijden een artikeltje waarin hij wetenschappelijk of technologisch nieuws inzichtelijk maakte, of de wetenschappelijke/technologische aspecten van een actualiteit uiteenzette. Voor deze opdracht doe je hetzelfde, maar nu over je thema uit de neerlandistiek of communicatiewetenschappen, op basis van je set artikelen.

Opdracht:
Schrijf in de stijl en volgens het stramien van de bijgevoegde voorbeeld-artikelen van Steketee een artikeltje van 500 woorden (+/- 5%) over een onderzoek in één van de bestudeerde artikelen in je thema-pakket. Misschien heb je na globale lezing de stijl van Steketee al te pakken en kun je onmiddellijk aan de slag. Niettemin is het raadzaam om ook na te gaan wat de kenmerken van die stijl precies zijn; zo kun je ze benoemen en krijg je grip op de technieken. Doelgroep: kinderen van 12 / 13 jaar.

Algemene aanwijzingen:
- Oriënteer je op de voorbeelden in de bijlage (driemaalSteketee.pdf) bij deze opdracht.
- Oriënteer je op (leesgedrag, interesses van) de doelgroep.
- Lees ter verdere oriëntatie de pagina's over stijltechnieken op www.rug.nl/noordster, het hoofdstuk over 'leesgemak' in de Schrijfwijzer van Renkema, en orienteer je op de thema's in het boek 'Formuleren' van Onrust e.a..
- Controleer vóór inlevering de richtlijnen voor het aanleveren van teksten in de studiehandleiding.
- Zet duidelijk bovenaan het artikel: 'concept'.
- Deadline: zie de studiehandleiding en het collegeoverzicht.
- Bij de beoordeling van deze opdracht zullen de docenten nagaan in hoeverre de tekst doelgroep-gericht is geschreven en in hoeverre je er in geslaagd bent de typische 'Steketee-stijl en het stramien van deze artikeltjes te imiteren.
APPENDIX 2: FEEDBACK FORM GRADE 7 STUDENTS (IN DUTCH)

Beste leerling,

Bedankt dat je de studenten wilt helpen om hun tekst te verbeteren!

In de enquête wordt er naar een cijfer gevraagd en of je het wilt uitleggen. Deze uitleg is erg belangrijk: op deze manier weet de student waarom je iets vindt.

1. Wat is de voornaam en de achternaam van de student?

2. Wat is het studentnummer van de student?

3. Wat vind je van het onderwerp?
   Ik vind het onderwerp (omcirkel het juiste antwoord):
   Helemaal niet leuk 1 2 3 4 5 Heel erg leuk

4. Leg kort uit waarom je dit vindt

5. Wat vind je van de titel?
   Ik vind de titel (omcirkel het juiste antwoord):
   Helemaal niet goed 1 2 3 4 5 Heel erg goed

6. Leg kort uit waarom je dit vindt

7. Wat vind je van het begin van de tekst (eerste alinea van de tekst)?
   Ik vind het begin van de tekst (omcirkel het juiste antwoord):
   Helemaal niet goed 1 2 3 4 5 Heel erg goed

8. Leg kort uit waarom je dit vindt

9. Wat vind je van het slot van de tekst (de laatste alinea)?
   Ik vind het slot van de tekst (omcirkel het juiste antwoord):
   Helemaal niet goed 1 2 3 4 5 Heel erg goed

10. Leg kort uit waarom je dit vindt

11. Wat vind je van de opbouw (structuur) van de tekst?
   Ik vind de opbouw van de tekst (omcirkel het juiste antwoord):
   Helemaal niet duidelijk 1 2 3 4 5 Heel erg duidelijk

12. Leg kort uit waarom je dit vindt

13. Wat vind je goed aan de tekst? Leg kort uit waarom
14. Wat vind je minder goed aan de tekst? Leg kort uit waarom

15. Hoe grappig vind je de tekst?
Ik vind de tekst (omcirkel het juiste antwoord):
Helemaal niet grappig  1  2  3  4  5  Heel erg grappig

16. Leg kort uit waarom je dit vindt

17. Hoe geschikt vind je de tekst voor jou en je klasgenoten?
Ik vind de tekst (omcirkel het juiste antwoord):
Helemaal niet geschikt  1  2  3  4  5  Heel erg geschikt

18. Leg kort uit waarom je dit vindt

19. Hoe aantrekkelijk vind je de tekst voor jou en je klasgenoten?
Ik vind de tekst (omcirkel het juiste antwoord):
Helemaal niet aantrekkelijk  1  2  3  4  5  Heel erg aantrekkelijk

20. Leg kort uit waarom je dit vindt

21. Hoe begrijpelijk vind je de tekst voor jou en je klasgenoten?
Ik vind de tekst (omcirkel het juiste antwoord):
Helemaal niet begrijpelijk  1  2  3  4  5  Heel erg begrijpelijk

22. Leg kort uit waarom je dit vindt

23. Wat vind je van de verzorging (geen taal- en spelfouten) van de tekst?
Ik vind de tekst (omcirkel het juiste antwoord):
Helemaal niet verzorgd  1  2  3  4  5  Heel erg verzorgd

24. Leg kort uit waarom je dit vindt

25. Zou je meer teksten willen lezen van deze studentschrijver? (omcirkel het juiste antwoord)
Ja  Nee

26. Wat voor cijfer geef je de tekst? (omcirkel het juiste antwoord)
1  2  3  4  5  6  7  8  9  10

Heel erg bedankt voor je hulp! 😊
APPENDIX 3: INSTRUCTION FOR WRITING A PEER REVIEW (IN DUTCH)

Schrijf twee korte peerreviews bij de opdrachtuitwerkingen van twee groepsgenoten. Zie de studiehandleiding voor informatie over het rolatiesysteem bij de toewijzing van teksten en deadlines. Omvang: 250 wrd (+/- 5 %). Vermeld bovenaan het review duidelijk: van wie de feedback afkomstig is en voor wie de feedback is bedoeld. Plaats je reviews voor aanvang van het college op Nestor en neem en print van zowel je samenvattingen als van de door jou geschreven reviews mee naar het college. En: neem zowel de DOOR jou geschreven reviews als de VOOR jou geschreven reviews op in je dossier. Let op: het is NIET de bedoeling met de 'edit'-functie van 'word' herschrijfsuggesties IN de tekst te geven. Zie hieronder een aantal suggesties voor constructieve feedback.

Je kunt constructieve feedback geven door:

- Aan te geven wat je de sterke punten van de tekst vindt, en op welke punten de tekst naar jouw idee sterker kan (geef bijvoorbeeld aan welke woorden, formuleringen of passages je aanspreken, indruk maken of juist ergernis wekken);
- Daarbij de niveaus te benoemen waarop jouw commentaar is gericht: inhoud, structuur, formulering/stijl, brongebruik, lay-out (en geef aan op welk gedeelte uit de tekst jouw commentaar betrekking heeft: welke pagina, paragraaf, alinea, zin binnen die alinea);
- Te proberen in eigen woorden te formuleren wat de onderliggende gedachte of eigenlijke boodschap van de tekst is (lukt dat?);
- Enkele suggesties ter verbetering te geven (pas op: het is niet de bedoeling dat je de tekst gaat redigeren).
Een milieuvriendelijke boomkor

MENNO STEKETEE


Duizenden jaren geleden bedacht iemand dat je op het gras voor je grot graan kon planten en schapen houden. Van dat graan kon je meel malen en brood bakken. Die schapen kon je melken en later opeten en van de wol kon je mooie truinen maken. Daarmee was de landbouw uitgevonden.

Honderden jaren geleden bedacht iemand dat je met behulp van stoommachines snel, meer graan kon malen, meer broden bakken en meer truinen maken. Toen was het industriële tijdperk aangebroken.

En iets meer dan tien jaar geleden schoot een slimme licht te binnen dat je al die broden en truinen via www.brodenentruien.com kon verkopen. Toen hadden we de informatiemaatschappij.

Het gekke is nu dat de meeste vissers zich na al die tijd nog steeds als die aapmensen gedragen. Het zijn nog altijd jagers-en-verzamelaars: ze varen met hun vissersvissers uit, gooien hun netten uit en vangen wilde vissen.

In de Noordzee doen ze dat zo grondig dat er niet zo veel vis meer over is. De grote schuldige van die overbevissing is de boomkor. Dat klinkt als een soort bosvogel - „op het menu van de schuwe boomkor staan vooral toren en kevers" - maar is een verwoestend type vierslacht.

Zo'n boomkor ploegt namelijk met zware kettingen over de zeebodem zodat de vissen uit het zand in het net worden gejaagd. Maar de zeebodem wordt daardoor ook verschrikkelijk omgewoeld. Daarom is een groot deel van de bodem van de Noordzee een soort woestijn geworden. Op de boerderij trekken ze toch ook niet twee keer per jaar een kename over het grasveld met schapen. Nou dan.

Daarom is er sinds kort de pulskor. Dat klinkt als een planeet uit de nieuwste Star Wars - „Zork, zet koers naar Pulskor" - maar is een milieuvriendelijk soort boomkor. Die gebruikt namelijk geen kettingen maar schrikdraad waar stroomstootjes doorheen schieten. Hierdoor schrikken de vissen het zand uit en komen alsnoog in het net terecht. Als klap op de vuurpijl kost het ook nog minder brandstof om de pulskor te slepen.

Op dit moment heeft nog maar één vissersvissers de pulskor op proef, de UK 153. Dat is nog veel te weinig. Maar of alle vissers graag willen werken met de pulskor is de vraag. De minister van vissen zou deze pulskor dus moeten verplichten. En voor veel vissers is dat anathema. Dat klinkt als een exotische struik - „de boomkor scharrelt onder de anathema" - maar dat betekent gewoon dat die vissers er helemaal geen zin in hebben.

Onvrijwillig transseksueel

MENNO STEKETEE

Sommige mannen denken dat ze een vrouw zijn, een vrouw opgesloten in het lichaam van een man. Ze gedragen zich daarom zoals ze denken dat een vrouw zich gedraagt. Ze treuzelen bijvoorbeeld bij het verlaten van het huis.

Ze lachen met een hoger stemgeluid, kunnen geen kaartiezen maar onthouden wel de verjaardagdatum van Jan en alleman. En ze kijken in iedere spiegelende winkelruit of hun haar nog wel goed zit.


Zulke mensen heten transseksueel. Transseksueelen kunnen geen kinderen maken.
Een zilveren randje

MENNO STEKETEE

Er is iets geks aan de hand met die orkanen die de hele tijd op Amerika afrazen. Die heten Katrina of Rita. En nou weer Wilma. Dat zijn meisjesnamen. Dat is helemaal niet leuk als je Katrina, Rita of Wilma heet.

De helft van de orkanen wordt naar jongens vernoemd, zoals Stan en Larry, maar het is net of die iets minder hard waaïen. Gek. Ze zouden beter een lange lijst met namen van dieren kunnen nemen waarvan iedereen weet dat ze gemeen zijn. „Texas op de vlucht voor Piranha.” Of: „Paalworm kost 84 miljard dollar.” Dan hoeft niemand zich beledigd te voelen.

In Europa is er ook iemand niet helemaal snik geweest bij het benoemen van winden. Hier heeft iemand het in zijn bolle kop gehaald ze naar Duitse auto’s te vernoemen: Sirocco, Passat, Bora.

Er zijn de laatste tijd veel natuur rampes. Dan is er weer een vloedgolf die tropische vissersdorpjes met de hele inhoud wegspoelt. Dan heeft de aarde wel weer ergens hele steden in elkaar. En er zijn dus veel orkanen. Toch zitten er heel soms ook goede kanten aan rampen. In veel landen bestaan daar zelfs gezegdes over. Iedere wolk heeft een zilveren randje, zeggen ze in Engeland. Niets is zo slecht of er komt wel iets goeds van, zeggen ze in Spanje en Rusland. Of, zoals een Amsterdamse filosoof ooit zei: „Ieder nadeel hept se voordeel.”
Beoordelingsformulier

Jouw naam:

Schrijf hier het cijfer dat bovenaan de tekst staat:

Titel
- Ik zou de tekst niet gaan lezen als ik de titel zie.
- Ik zou de tekst zeker gaan lezen als ik de titel zie.
- De titel is niet geschikt voor een artikel voor een kinderpagina van een groot landelijk dagblad.
- De titel is heel geschikt voor een artikel voor een kinderpagina van een groot landelijk dagblad.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

Leg je cijfer kort uit:

Inleiding
- Na het lezen van de inleiding wil ik niet doorlezen.
- Na het lezen van de inleiding wil ik heel graag doorlezen.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

Leg je cijfer kort uit:

Slot
- Niet aantrekkelijk/afsluitend slot.
- Heel aantrekkelijk/afsluitend slot.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

Leg je cijfer kort uit:

Publieksgerichtheid
- De inhoud is niet interessant/geschikt voor kinderen van mijn leeftijd.
- De inhoud is heel interessant/geschikt voor kinderen van mijn leeftijd.
- De schrijver gebruikt voorbeelden/vergelijkingen die niet leuk zijn voor kinderen van mijn leeftijd
- De schrijver gebruikt voorbeelden/vergelijkingen die heel leuk zijn voor kinderen van mijn leeftijd

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

Leg je cijfer kort uit:

Publieksgerichtheid
- Het taalgebruik van de tekst is niet
- Het taalgebruik in de tekst is heel geschikt voor
**geschikt voor kinderen van mijn leeftijd**

- Ik voel me als lezer niet serieus genomen
- Ik voel me als lezer serieus genomen.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

**Leg je cijfer kort uit:**

**Structuur**

- De verbanden tussen de alinea’s zijn niet duidelijk en logisch
- De verbanden tussen de alinea’s zijn heel duidelijk en logisch

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

**Leg je cijfer kort uit:**

**Stijl**

- De schrijver heeft geen levendige/aantrekkelijke/humoristische stijl
- De schrijver heeft een heel levendige/aantrekkelijke/humoristische stijl

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

**Leg je cijfer kort uit:**

**Verzorging**

- In de tekst staan veel spel- en taalfouten
- In de tekst staan geen spel- en taalfouten
- In de tekst staan geen fouten met punten en komma’s
- In de tekst staan geen fouten met punten en komma’s

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

**Leg je cijfer kort uit:**

**Van wetenschappelijke tekst naar kinderpagina-artikel**

- De tekst lijkt meer voor wetenschappers geschreven dan voor kinderen.
- De tekst is echt voor kinderen geschreven

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
</table>

**Leg je cijfer kort uit:**
Imitatie
- De tekst van Menno Steketee is niet geïmiteerd.
- De zinnen en alinea’s zijn niet kort
- Er zit geen humor in de tekst
- De tekst heeft niet de structuur van de teksten van Menno Steketee

-De tekst van Menno Steketee is heel goed geïmiteerd.
- De zinnen en alinea’s zijn kort
- Er zit veel humor in de tekst
- De tekst heeft de structuur van de tekst van Menno Steketee

1 2 3 4 5

Leg je cijfer kort uit:

Eindoordeel:

1 2 3 4 5 6 7 8 9 10
### APPENDIX 6: CODEBOOK SPSS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable in SPSS</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of the text</td>
<td>Textnumber</td>
<td>Text 1-20</td>
</tr>
<tr>
<td>Number of feedback comments by peers</td>
<td>Number_peerfeedback</td>
<td>Total number of received feedback comments: peers</td>
</tr>
<tr>
<td>Number of feedback comments by G7 students</td>
<td>Number_G7_feedback</td>
<td>Total number of received feedback comments: G7</td>
</tr>
<tr>
<td>Total number of feedback comments: G7 and P</td>
<td>Total_number_feedback</td>
<td>Total number of received feedback comments: P and G7</td>
</tr>
<tr>
<td>Number of peer feedback comments implemented</td>
<td>Implementing_peerfeedback</td>
<td>Total number of implemented peer feedback</td>
</tr>
<tr>
<td>Mean of the peer feedback implementation</td>
<td>Average_peerfeedback_implementation</td>
<td>Mean of the number of implemented peer feedback</td>
</tr>
<tr>
<td>Numbers of G7 feedback comments implemented</td>
<td>Implementing_G7_feedback</td>
<td>Total number of G7 feedback comments implemented</td>
</tr>
<tr>
<td>Mean of the G7 feedback implementation</td>
<td>Average_G7_feedback_implementation</td>
<td>Mean of the number of implemented G7 feedback</td>
</tr>
<tr>
<td>Total number of feedback implemented: G7 and P</td>
<td>Implemented_total</td>
<td>Total number of feedback comments implemented: G7 and P</td>
</tr>
<tr>
<td>Mean of total implemented feedback</td>
<td>Average_implemented</td>
<td>Mean of the total number of feedback comments implemented: G7 and P</td>
</tr>
<tr>
<td>More implementation peer feedback or G7 feedback</td>
<td>Implementing_which_feedback_more</td>
<td>0= same, 1= peer, 2= G7</td>
</tr>
<tr>
<td>Number of changes based on feedback comments</td>
<td>Change_based_on_feedback</td>
<td>0= not based on feedback, 1= based on feedback</td>
</tr>
<tr>
<td>Changing more based on one feedback comment</td>
<td>More_changes_one_feedback_comment</td>
<td>0 = only one change per feedback comment, 1= More changes per feedback comment</td>
</tr>
<tr>
<td>Feedback comment that initiated more than one change: G7, P or both</td>
<td>Kind_of_feedback_based_on_more_changes</td>
<td>1 = peer, 2= G7, 3= both</td>
</tr>
<tr>
<td>Does the writer understand the feedback that initiated the change</td>
<td>Understanding_feedback</td>
<td>0 = does not understand the feedback, 1= understands feedback</td>
</tr>
<tr>
<td>Does the writer agree with the feedback that initiated the change</td>
<td>Agreement_feedback</td>
<td>0 = does not agree with the feedback, 2= agrees with the feedback</td>
</tr>
<tr>
<td>Scope of the revision</td>
<td>Scope_revision</td>
<td>1= surface, 2= word, 3= sentence, 4= paragraph, 5= text as a whole</td>
</tr>
<tr>
<td>Proportion of surface scope</td>
<td>Scope_proportion_surface</td>
<td>Proportion</td>
</tr>
<tr>
<td>Proportion of word scope</td>
<td>Scope_proportion_word</td>
<td>Proportion</td>
</tr>
</tbody>
</table>
Proportion of sentence scope  Scope_proportion_sentence  Proportion
Proportion of paragraph scope  Scope_proportion_paragraph  Proportion
Proportion of text scope  Scope_proportion_text  Proportion
Does the revision alter the meaning of the text  Revision_meaning_altering  Proportion

0= formal changes,
1= Meaning preserving_all (micro and macro),
2= meaning altering affects concepts,
3= meaning altering affects summary

Proportion of formal meaning preserving  Proportion_meaning_formal  Proportion
Proportion of meaning preserving micro and macro  Proportion_meaning_preser
Proportion of meaning altering micro (concepts)  Proportion
Proportion of meaning altering macro (summary)  Proportion
How many changes did the student make  Changes  Proportion
What kind of revision is it  Kind_of_revision
1= addition, 2= deletion,
3= substitution, 4= permutation,
5= distribution, 6= consolidation

Proportion of the operation addition  Proportion_addition  Proportion
Proportion of the operation deletion  Proportion_deletion  Proportion
Proportion of the operation substitution  Proportion_substitution  Proportion
Proportion of the operation permutation  Proportion_permutation  Proportion
Proportion of the operation distribution  Proportion_distribution  Proportion
Proportion of the operation consolidation  Proportion_consolidation  Proportion

Is the local change an improvement or not  Local_changes_improvements
_or_not
_1 = deterioration, 1 = improvement,
0 = no change

Proportion of no change  Proportion_improvement_same  Proportion
Proportion of improvement  Proportion_improvement  Proportion
Proportion of deterioration  Proportion_deterioriation  Proportion

How many people provided feedback which initiated the revision  Number_of_feedback_total  Number feedback providers
suggesting this change

What is the average end judgement of all the raters  End_judgement_concept  End judgement on a 1 - 10 scale
on the concept text
What is the average end judgement of all the raters on the final text
What is the difference in average grade between the concept and end text
What is the average end judgement of all the raters on the concept text
What is the average end judgement of all the raters on the final text
What is the difference in average grade between the concept and end text
The difference in value between the concept version and the end version

<table>
<thead>
<tr>
<th>Person who assessed the text</th>
<th>Pers.Nr</th>
<th>initials person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviewer who assessed the text: expert rater or grade 7 student</td>
<td>Type_reviewer</td>
<td>P= expert rater, G7= grade 7 student</td>
</tr>
<tr>
<td>Sex of the reviewer: male or female</td>
<td>Sex_reviewer</td>
<td>0 = female, 1 = male</td>
</tr>
<tr>
<td>Concept or final version of the text</td>
<td>Concept_eind</td>
<td>0 = concept, 1 = final version</td>
</tr>
<tr>
<td>Sex of the writer</td>
<td>Geslacht_schrijver</td>
<td>0 = female, 1 = male</td>
</tr>
<tr>
<td>Opinion on the title by the reviewer</td>
<td>Title</td>
<td>Likert scale: 1-5</td>
</tr>
<tr>
<td>Presence of title</td>
<td>No_title</td>
<td>1= not present, 0= present</td>
</tr>
<tr>
<td>Opinion on the introduction by the reviewer</td>
<td>Introduction</td>
<td>Likert scale: 1-5</td>
</tr>
<tr>
<td>Opinion on the end of the text by the reviewer</td>
<td>Ending</td>
<td>Likert scale: 1-5</td>
</tr>
<tr>
<td>Opinion on the content for this particular audience by the reviewer</td>
<td>Audience_content</td>
<td>Likert scale: 1-5</td>
</tr>
<tr>
<td>Opinion on the language for this particular audience by the reviewer</td>
<td>Audience_language</td>
<td>Likert scale: 1-5</td>
</tr>
<tr>
<td>Opinion on the structure of the text by the reviewer</td>
<td>Structure</td>
<td>Likert scale: 1-5</td>
</tr>
<tr>
<td>Opinion on the style of the text by the reviewer</td>
<td>Style</td>
<td>Likert scale: 1-5</td>
</tr>
</tbody>
</table>
Opinion on spelling and punctuation by the reviewer

Polish_spellingerror_punctuation
Likert scale: 1-5

Opinion on layout by the reviewer

Polish_layout
Likert scale: 1-5

Opinion on the transformation of a science article to a children's article by the reviewer

Translation_science_children
Likert scale: 1-5

Opinion on the imitation of Menno Steketee by the reviewer

Imitation
Likert scale: 1-5

Grade by the student for the usefulness of the peer feedback

Cijfer_brukbaarheid_peer
1= not useful at all, 2= not useful, 3= neither useful nor not useful, 4= useful, 5= very useful

Grade by the student for the usefulness of the G7 feedback

Cijfer_brukbaarheid_leerling
1= not useful at all, 2= not useful, 3= neither useful nor not useful, 4= useful, 5= very useful

Difference between peer and G7 student in usefulness of the feedback

Difference_between_peer_G7_usefulness
1= peer feedback more useful, 2= G7 feedback more useful

Grade for agreement with peer feedback

Cijfer_inkunnenvinden_peer
1= no agreement at all, 2= no agreement, 3= neither agree, nor disagree, 4= agree, 5= agree a lot

Grade for agreement with G7 feedback

Cijfer_inkunnenvinden_leerling
1= no agreement at all, 2= no agreement, 3= neither agree, nor disagree, 4= agree, 5= agree a lot

Difference between peer and G7 student in agreement with the feedback

Difference_between_peer_G7_agreement
1= peer feedback more useful, 2= G7 feedback more useful

Number of feedback comments per text by a peer

Implementation_comments_P
Number of feedback comments per text by a peer

Number of feedback comments from a peer that have been implemented

Percentage_comments_implemented_P
Percentage of the number of feedback comments from a peer that have been implemented

Number of feedback comments per text by a grade 7 student

Number_of_comments_G7
Number of feedback comments per text by a G7 student

Number of feedback comments from a G7 student that have been implemented

Implementation_comments_G7
Number of feedback comments from G7 students that have been implemented

Percentage of the number of feedback comments G7 student

Percentage_comments_implemented_G7
Percentage
<table>
<thead>
<tr>
<th>Description</th>
<th>Formula</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perception of receiving and writing feedback</td>
<td>Perception_feedback</td>
<td>1= Very negative perception, 2= Negative perception, 3= Neutral, 4= Positive perception, 5= very positive perception</td>
</tr>
<tr>
<td>Mean of all analytical items summed up</td>
<td>All_items_scale_mean</td>
<td>Mean of all items on analytical scale added up</td>
</tr>
<tr>
<td>Mean of agreement and usefulness of the feedback from peers</td>
<td>Agreement_and_usefulness_peer</td>
<td>Mean of the agreement and usefulness feedback peers</td>
</tr>
<tr>
<td>Mean of agreement and usefulness of the feedback from G7 Students</td>
<td>Agreement_and_usefulness_G7Student</td>
<td>Mean of the agreement and usefulness feedback G7 students</td>
</tr>
<tr>
<td>Mean of agreement and usefulness of the feedback from G7 Students and peers</td>
<td>Agreement_and_usefulness_peer_G7Student</td>
<td>Mean of agreement and usefulness feedback from G7 students and peers</td>
</tr>
<tr>
<td>End judgement on either the concept text or the final text</td>
<td>Eindoordeel</td>
<td>End judgement on a 1 - 10 scale</td>
</tr>
</tbody>
</table>