# Assessing writing ability in primary education

*On the evaluation of text quality and text complexity*

Hiske Feenstra

**ASSESSING WRITING ABILITY IN PRIMARY EDUCATION**

ON THE EVALUATION OF TEXT QUALITY AND TEXT COMPLEXITY

**Hiske Feenstra**

**Graduation committee**

| | | |
|---|---|---|
| Chairman | prof. dr. ir. A.J. Mouthaan | University of Twente |
| Promotores | prof. dr. ir. T.J.H.M. Eggen | University of Twente |
| | prof. dr. T.J.M. Sanders | Utrecht University |
| Members | prof. dr. H.H. van den Bergh | Utrecht University |
| | prof. dr. A.P.J. van den Bosch | Radboud University Nijmegen |
| | prof. dr. C.M. de Glopper | University of Groningen |
| | prof. dr. T.W.C. Huibers | University of Twente |
| | prof. dr. ir. B.P. Veldkamp | University of Twente |

ISBN: 978-90-365-3725-4
DOI: 10.3990/1.9789036537254
Printed by Ipskamp Drukkers, Enschede
Cover designed by Studio Het Mes, Den Haag
© 2014, H.M. Feenstra. All rights reserved.

*This research was supported by Cito, Institute for Educational Measurement.*

**ASSESSING WRITING ABILITY IN PRIMARY EDUCATION**

ON THE EVALUATION OF TEXT QUALITY AND TEXT COMPLEXITY

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 5 september 2014 om 12.45 uur

door

Hiske Marianne Feenstra

geboren op 24 augustus 1981

te Groningen

This dissertation has been approved by

| | |
|---|---|
| prof. dr. ir. T.J.H.M. Eggen | University of Twente |
| prof. dr. T.J.M. Sanders | Utrecht University |

## Table of contents

# 1 Introduction



## 1.1 Assessing writing

A written text is the product of a complex set of activities. Producing a text first of all involves the mental processes of generating, organising, and structuring ideas, and translating thoughts into words. Then specific motor skills are needed to produce letters in either handwriting or typewriting. Lastly, writing involves monitoring and editing the text produced. Together, these activities constitute the concept of writing ability.

The ability to produce a well-written text is an important skill to master in one's educational career and to apply in one's professional and everyday life. Whereas the production of spoken text occurs spontaneously in young children, explicit instruction is needed to produce written language. Compared to the acquisition of speech, learning to produce written language is a scholastic skill that language learners do not develop spontaneously. Rather, thorough instruction is needed to develop this skill—in contrast to producing spoken language, for which no explicit instruction is needed. However, defining the precise skills needed to construct a written text is challenging. Hence, different branches within writing research are dedicated to unravelling the construct of writing ability, improving the teaching of writing, and refining methods to assess writing quality in order to evaluate teaching success.

### 1.1.1 The construct of writing ability

The present ideas on the construct of writing ability are the result of several decades of research on writing. In the 1980s, the focus of writing research shifted from studying the effects of employing different writing didactics to building theories on the writing process (Schoonen & De Glopper, 1992). Cognitive psychologists employed so-called 'think aloud protocols' to model the cognitive process of writing a text. An influential model proposed by Flower and Hayes (1981) consisted of three components: the writer's long-term memory (i.e., knowledge on topic and audience, and writing plans), the different writing processes (i.e., planning, translating, reviewing, and monitoring), and the task environment (i.e., the rhetorical problem). Over the years, the model has evolved under the influence of writing research, resulting in a model that comprises a resource level (covering, e.g., working memory and long-term memory), a processing level (covering the task environment and the actual writing processes), and a control level in which the writing goal is set and controlled.

Based on theories on the writing process by Flower and Hayes (1981) and Hayes (2012), Figure 1 provides a schematic representation of the different components of the writing process and their interaction.

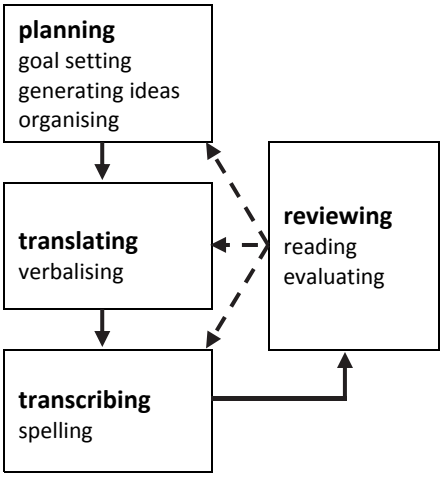



*Figure 1.* A schematic representation of the writing process.
(based on Flower & Hayes, 1981; Hayes, 2012)

Apart from the aforementioned shift in focus within writing research, developing ideas on the processes underlying writing influenced the practice of writing instruction as well. In the last decades of the twentieth century, a transition from *product*-oriented to *process*-oriented writing instruction took place, fuelled by research on the modelling of writing. In line with this, a move away from emphasising mechanics (e.g., spelling and grammar) towards purpose (e.g., context, tone, audience, and conventions) has been observed in recent years (Hardison & Sackett, 2008; Rijlaarsdam et al., 2012). Eventually, this transition has led to a classroom practice in which evaluation of the writing process is employed to diagnose strengths and weaknesses in writing and to use this information within writing instruction (Rijlaarsdam et al., 2012).

### 1.1.2 Assessing writing ability

To be able to check whether (changes in) didactic approaches to the teaching of writing are effective, the quality of the results of the writing process (i.e. writing products) has to be evaluated. Hence, the assessment of writing quality is a key element in the practice of writing instruction, as well as in research on didactics. Assessing writing ability can be described as a chain of inferences; many steps are needed to construct a score that represents this productive language skill. The process from writing task to writing score is represented schematically in Figure 2.

First, a writer is assigned a writing task, which is selected out of the domain of all possible tasks—differing in topic, intended audience, text genre, communicative goal, etc.

Then the writer engages in the complex practice of producing a text, a process that is influenced by the writer's specific characteristics, such as personality and motivation. The result of this process is a writing product, that is, a performance that represents the writer's ability. Next, one or more raters evaluate this performance by applying certain rating procedures, the result of which is a text quality rating. The last step is to transform this evaluation into a score on writing ability. Ideally, this procedure would lead to scores that are comparable across different tasks, rating procedures, and rater(s). However, in practice, large unwanted effects of these variables are found (cf. Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012).

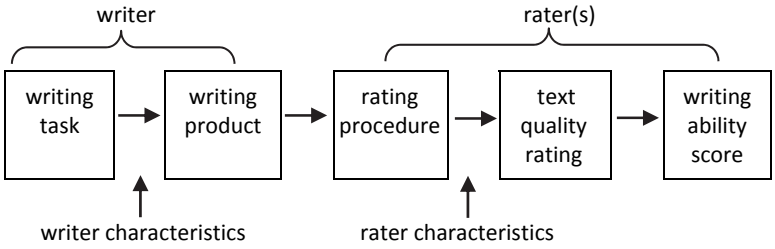



*Figure 2.* From writing task to writing score.

Over the past decades, a large body of studies has been conducted to explore different approaches to the assessment of writing, taking into account reliability, validity, and generalisability. The reliability of an assessment concerns the accuracy of scores produced by the assessment. Within writing research, this accuracy is dependent on the stability of scores assigned by raters, that is, the agreement *amongst* different writers and *within* a single rater over time (inter-rater and intra-rater agreement, respectively). The validity of an assessment is described as "the degree to which evidence and theory support the interpretation of test scores entailed by their proposed use" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

In writing research, validity studies can concern both the degree to which a writing task is an authentic representation of an actual writing situation and the degree to which a rating of the writing product truly represents the quality of the performance under consideration. Generalisability combines issues of both reliability and validity, by evaluating the degree to which a score that resulted from a given test design is generalisable across other designs (i.e., different tasks, raters, and rating procedures). Over the years, the focus of studies on writing assessment has shifted from the comparison of direct versus indirect measurement approaches to specific issues concerning the reliability and validity of essay scores, for instance, by studying rater effects and evaluating generalisability.

Direct versus indirect assessment

A so-called *direct* assessment of writing ability by means of an essay assignment is generally considered more adequate than the administration of multiple-choice tests as (indirect) measures of writing ability, given that no actual writing is required when answering multiple-choice questions (Cooper, 1984). However, the reliability and validity of these essay assignments can be impaired because of several factors influencing test scores on writing ability. In other words, raters are known to disagree on writing scores, and the scores given can reflect irrelevant (invalid) factors. In the twentieth century, these facts prompted an extensive debate about the validity and reliability of the direct and indirect approaches, in which the advantages and disadvantages of both methods were discussed. Table 1 provides a summary of the characteristics of both methods, based on Cooper (1984).

First, an obvious benefit of direct writing assessments is their high face validity. Regardless of their actual appropriateness, assessments in which the candidates are prompted to produce a written text come across as valid because of their appearance. For indirect assessment, face validity is impaired since no actual writing is required. On the other hand, indirect measures mostly comprise multiple-choice questions, meaning that they can be objectively scored, and the scores are not influenced by rater effects. Together with the effects of task characteristics, these rater effects can impair both the reliability and validity of direct assessments. Lastly, direct assessments are thought to have a positive effect on the teaching of writing, not only by conveying the message that writing is important, but also because of the possibility of providing a writer with feedback on the writing product, in order to improve future performances. While indirect assessment does offer the opportunity to assess certain specific elements of writing, such as revision skills, objectively scored tests can only partly cover the domain of writing, and possibly give rise to a negative wash-back effect on the teaching of writing. In other words, employing an indirect assessment to evaluate the writing skill might lead to a practice in which the teaching of writing is about preparing for the specific topics addressed in this indirect measurement instead of learning to write.

Table 1. *Characteristics of Direct and Indirect Assessment* (Cooper, 1984)

| **Direct assessment** | **Indirect assessment** |
|---|---|
| + High (face) validity | + Possibility to assess specific elements |
| + Positive effect on teaching of writing (e.g., formative assessment) | + Scored objectively and economically |
| − Task/rater effects are a source of unwanted variance | − No writing required (impaired face validity) |
| | − Covers limited (peripheral) area within domain of writing skill |
| | − Negative wash-back effect on teaching of writing |

Reliability and validity of essay scores

In recent years, the focus of writing research has moved away from the use of indirect assessment, in part because this method implies the avoidance of those features of writing on which raters disagree. Instead, efforts have been made to ensure that raters can deal with the inevitable problems of rating. This stance was also argued by Clauser (2000), who noted that strategies to increase objectivity could be a threat to validity. Hence, efforts to raise agreement between raters should not lead to avoidance of those features raters disagree on. Rather, a clear focus on these features should be accomplished. As Clauser stated: "Alternatively, if the focus is shifted away from these features—toward features that are more easily quantified—increased reliability could lead to decreased validity." (Clauser, 2000, p. 313). When attempting to improve the reliability and validity of essay scores, three variables are of particular importance: the properties of the specific *task* on the basis of which an essay is to be written, the *rating procedure* that is used to evaluate the essay, and the characteristics of the *rater(s)* assigned to perform this evaluation.

Research has shown that the specific writing task assigned to evaluate writing ability influences the scores obtained. Van den Bergh (1988) found that task-related variance exceeded the proportion of variance related to both the difference in ability amongst writers and the disagreement amongst raters. In line with these results, Rijlaarsdam et al. (2012) stated that scores for different writing tasks often show only moderate or even low correlations. These results imply that writing scores are largely dependent on the specific task assigned, and hence, no reliable and valid assumptions on the writer's ability across different tasks can be made on the basis of the performance on one single writing task. Instead, several tasks are needed to be able to generalise the scores obtained to the construct of writing ability.

Apart from the writing task, the rating procedure is a possible source of (unwanted) variance amongst writers. An array of different rating procedures is available, varying in the amount and type of information they provide on the writing performance. Wolcott and Legg (1998) provided an overview of the advantages and disadvantages per method, which are summarised in Table 2. As for holistic scoring, this method can provide reliable scores in a relatively short amount of time; on the other hand, it does not provide information on the performance on different aspects of writing ability. Analytical scoring methods are suited to provide more extensive information on a writing performance, meaning that the method can also be applied for formative assessment purposes. However, detailed scoring schemes have to be constructed, and the scoring process can be time consuming.

By way of an adaptation of holistic scoring, the 'primary trait' procedure offers a method in which writing products are evaluated solely with respect to their communicative effectiveness. Within this method, the degree to which the writing product achieves the specific goal stated for this assignment determines the score given to the product. While this holistic method is efficient and specifically concerns the essence of writing a text, it does not offer detailed information on the performance. In order to offer raters fixed reference points

during the process of rating several essays, the practice of providing raters with exemplar essays of performances of different qualities has been developed. However, this method is difficult to construct for individual teachers since it requires a meticulous and informed selection of the exemplars.

When considering the characteristics of both holistic and analytical scoring, their advantages and disadvantages are determined by the intended use of the test scores. With respect to a large-scale assessment, practical downsides such as the lack of efficiency are less problematic when compared to classroom assessment. Typically, only a sample of students per class will participate in a large-scale type of assessment, and a large pool of raters will be employed to extensively evaluate the writing products.

Table 2. *Characteristics of Different Rating Procedures* (Wolcott & Legg, 1998)

| Rating procedure | Characteristics |
|---|---|
| **Holistic**<br>(impression of the whole essay in one rating) | • Efficient<br>• No information on aspects of writing performance |
| **Analytical**<br>(rating per aspect, clearly defining scores per aspect of writing) | • Time consuming<br>• Elaborate information on different aspects of writing performance |
| **Primary trait**<br>(text is only judged on how well the text goal is achieved) | • Efficient<br>• Concerns essence of writing; less informative |
| **Exemplar essays**<br>(used as reference points) | • Careful selection of exemplars needed<br>• Offers fixed standard |

Lastly, a number of rater effects are identified in the literature on the assessment of writing. Rijlaarsdam et al. (2012) summed up a series of effects that have been observed when rating text quality and that can negatively affect the reliability and validity of the writing assessment. These effects are presented in Table 3, together with approaches to possibly overcome these effects. First, the fact that individual raters differ in their ideas on the relative importance of the elements of writing quality might lower the agreement amongst them and hence decrease reliability (*signific effect*). This effect can be minimised by providing raters with a clear instruction on the aspects to consider, and their relative weights or by asking them straightforward yes/no questions on specific features of the essay. Second, one specific feature of a performance can influence the overall rating (*halo effect*), which threatens both the agreement amongst raters and the validity of the assessment. Providing raters with exemplar essays as references or assessing different aspects of writing in separate rounds is likely to lower this effect.

Next, the order in which essays are presented to a rater is known to influence the scores (*sequence effect*), that is, the relative quality of a preceding essay affects the next performance to be rated. This effect should be accounted for by randomising the order of essays per rater, but it can also be lessened by offering exemplar essays as fixed reference points—as an alternative to comparing each essay to the preceding essay. A so-called *norm shift* occurs when a rater unknowingly adapts his or her norm to the overall quality of the essays. Again, providing raters with a standard by means of exemplar essays is likely to minimise this effect. Furthermore, an array of *rater characteristics* is likely to influence the manner in which raters evaluate writing performances. By assigning multiple raters to an essay, rater effects can be levelled out. Additionally, offering exemplar essays as a comparison is likely to improve agreement amongst raters. Lastly, a so-called *contamination effect* occurs when performance ratings are used to accomplish goals other than assessment, for example, rewarding or punishing a pupil for his or her behaviour in class. Anonymising the essays will prevent writers' characteristics (other than writing ability) from influencing the individual rating.

Table 3. *Rater Effects* (Rijlaarsdam et al., 2012)

| Rater effect | Description | Solution(s) |
|---|---|---|
| **Signific effect** | Raters differ in their conception of what elements are important | • Provide clear and stringent instructions |
| **Halo effect** | Irrelevant aspect of the performance influences the score on other aspects | • Assess different aspects in separate rounds<br>• Provide exemplars as fixed references |
| **Sequence effect** | Sequence in which essays are rated influences the rating | • Randomise sequence per rater<br>• Provide exemplars as fixed references |
| **Norm shift** | Rater unconsciously adapts norm to the quality of the essays | • Provide exemplars as fixed references |
| **Rater characteristics** | Raters have a lenient, central or strict tendency | • Assign multiple raters per essay<br>• Provide exemplar essays as fixed references |
| **Contamination effect** | Entanglements of interests, for example, when teachers are influenced by a pupil's classroom behaviour | • Anonymise essays |

## 1.2 Research goal and outline

### 1.2.1 Research goal

As indicated in the previous section, the assessment of writing is complicated by the multifaceted nature of writing on the one hand, and the difficulty of evaluating writing performances, on the other hand. Whereas a large body of research is available on the assessment of writing, many of the existing studies are targeted at the assessment of second language learners or are conducted within groups of more or less advanced writers, that is, pupils aged 12 years or older. Hence, relatively little is known about the evaluation of texts produced by novice writers and the specific issues that arise when evaluating their written products. In Figure 3, the writing performance of novice writers is illustrated.


*Figure 3*. Essays written by novice writers.
(aged 8 to 12, grades 3 through 5 in primary education)

The present study aims to explore how to design a writing assessment that is suited for assessing different aspects of writing ability in primary education in a valid and reliable manner. To do so, the traditional approach of evaluating text quality is adopted, as well as a more innovative approach, which explores the use of language technology to evaluate text complexity. Because text complexity features are objectively identifiable aspects of texts, this approach is expected to provide a writing score that is both stable and valid. That is, the

automated evaluation of complexity scores does not depend on the processing of this text by a certain rater; instead, it represents an inherent property of the evaluated text.

Research context

This study is performed within the context of the Dutch national assessment in primary education, which aims at evaluating the educational system by means of an assessment of writing ability within a sample of schools and pupils. In past cycles of the Dutch national assessment (Krom et al., 2004; Kuhlemeier, Van Til, Hemker, De Klijn, & Feenstra, 2013; Sijtstra, 1997; Zwarts, 1990), different approaches to the measurement of writing ability were adopted. In the first two cycles, essays were holistically scored following the primary trait principle; informative texts were evaluated according to their communicative value, and narrative texts were evaluated based on their entertainment value. Raters used a 0 to 200 scale, accompanied by exemplars representing a 100-point essay in which all relevant elements were present. Raters were given a method to add points when more elements were present to enhance the quality. If essential elements were missing, points were deducted (Sijtstra, 1997; Zwarts, 1990). However, scores produced via this procedure proved difficult to interpret, and the construction of a single ability scale for all writing assignments did not succeed, indicating problems with the representation of the construct of writing by the primary trait scores.

Because of the aforementioned problems when employing the primary trait procedure, an analytical evaluation of different aspects of writing was introduced in the subsequent cycle (Krom et al., 2004). In this method, the task of raters was reduced to answer simple yes/no questions on specified characteristics of the essays. To ensure comparability across cycles, this method was adopted in the next cycle as well (Kuhlemeier et al., 2013). However, the reliability and validity of this procedure were arguably impaired, since a relatively low agreement amongst raters was achieved, and no overall evaluation of the texts as a whole was provided—giving rise to the question of how to improve the assessment procedure and providing a rationale for the current study.

As described in the preceding sections, the various components of a writing assessment are potential sources of construct-irrelevant variance (Messick, 1989), for example, assessment method, writing task, rating procedure, and rater characteristics. Although recognised as unwanted effects that affect the generalisability of writing scores (cf. Schoonen, 2005; Van den Bergh et al., 2012), not all of these sources can be manipulated or eliminated easily. For example, task effects could be ruled out by dramatically increasing the number of tasks given to a student, and the effect of rater characteristics could be limited by assigning a large number of different raters to evaluate each essay. However, the feasibility of these methods is subject to the resources available, given the time and effort they would require of both students and raters. Instead, when altering the manner in which a performance or a performance rating is collected (i.e., adapting the task or the rating

procedure), the reliability and validity of a writing assessment are less dependent on the available means.

This study therefore focuses on altering (parts of) the *assessment method* as well as the *rating procedure* of the Dutch national assessment to improve the reliability and validity of the writing assessment. In Figure 4, specific elements of the writing product are related to a simplified model of the writing process (cf. Figure 1). In this representation, the conceptual process of *planning* what to write by generating and organising ideas is linked to the overall structure and content of a writing product (i.e., macro level). The relations amongst sentences and within the sentence structure, together with word choice and spelling (i.e., meso and micro level), reflect the process of *translating* these ideas into words. The action of *transcribing* these words onto paper or screen is reflected in the orthography of a written product (i.e., micro level). The *reviewing* and—if needed—editing of a text comprise a separate component within the writing process, the results of which can affect all essay characteristics. In this dissertation, different methods to evaluate writing products are examined, in order to separately assess novice writers' execution of all four parts of the writing process.

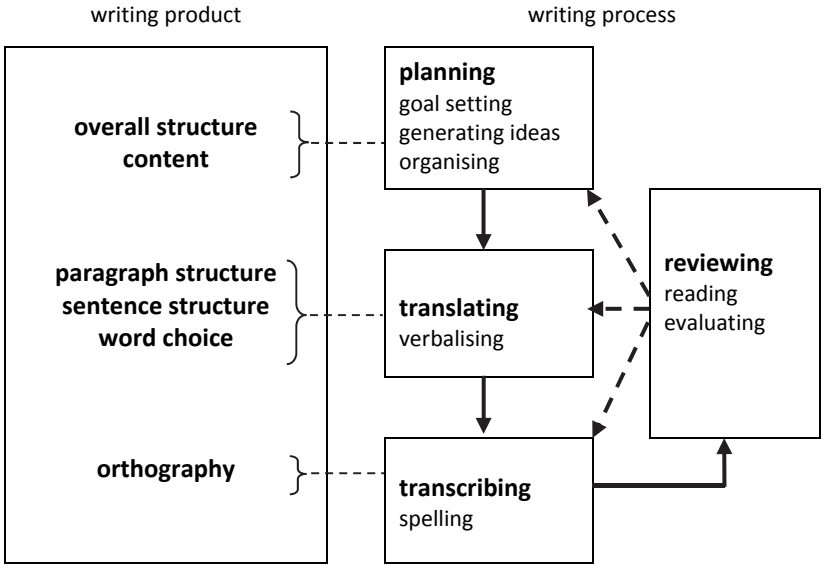



*Figure 4.* Elements of the writing product mapped to components of the writing process. (based on Flower & Hayes, 1981; Hayes, 2012; Van der Pool, 1995)

### 1.2.2 Research questions

The present study aims to explore different methods to improve the reliability and validity of the current Dutch national assessment of writing ability by answering the following question:

*How can the writing ability of novice writers be assessed reliably and validly?*

To answer this question, three different approaches to the assessment of writing are evaluated within the context of a large-scale assessment. First, a newly developed procedure, in which a rating scale with anchor essays is added to an analytical rating procedure, is evaluated. In Chapter 2, the inter-rater agreement, generalisability, and construct validity of this anchored analytical assessment (AAA) procedure is addressed by means of the following research questions:

*2.1 To what extent does the addition of anchor essays to an analytical evaluation provide higher inter-rater reliability when compared to a solely analytical procedure?*
*2.2 To what extent are writing scores obtained via the anchored analytical assessment generalisable across tasks and raters?*
*2.3 To what extent do writing scores obtained via the anchored analytical assessment provide evidence for construct validity?*

Adding a fixed reference point is expected to increase agreement amongst raters and provide a writing score that accurately reflects the overall quality of the essay, hence improving both reliability and validity. Furthermore, writing scores obtained via the AAA procedure are expected to be better generalisable across raters and tasks when compared to a solely analytical procedure.

Chapter 3 discusses the use of revision tests within an assessment of writing. First, the validity of an existing standardised, multiple-choice revision test is evaluated. Additionally, a constructed-response version of this test is piloted, and its validity and test characteristics are compared to the existing multiple-choice version. The following research questions are answered:

*3.1 To what extent does the content of the current multiple-choice revision test represent the content domain of revision ability?*
*3.2 To what extent do correlations with related constructs (reading, vocabulary, and writing) and non-related constructs (arithmetic) support the construct validity of the current multiple- choice revision test?*
*3.3 To what extent does the content of the piloted, constructed-response revision test represent the content domain of revision ability?*

*3.4 To what extent does the piloted, constructed-response revision test provide for equivalent test characteristics (i.e., reliability, difficulty, and discriminative power) when compared to the multiple-choice version?*

Evaluating revision ability within an assessment of writing is expected to increase the validity of the assessment, since it ensures coverage of an important part of the writing process. Taking both construct coverage and face validity into account, a combined multiple-choice and constructed-response revision test is expected to be best suited within an assessment of novice writers.

Lastly, in Chapter 4, the use of automated essay evaluation (AEE) is explored in order to gain knowledge on the relation between text complexity measures and writing ability, as well as the possibilities of using these measures as part of a writing assessment in primary education. In this exploratory study, the following research questions are addressed:

*4.1 What specific difficulties arise in analysing essays of novice writers? To what extent can these be overcome?*
*4.2 Which text complexity measures are suited to describe and evaluate the development in writing ability from mid-primary education (grade 3/4) to end-primary education (grade 5/6)?*
*4.3 To what extent are the selected measures valid indicators of writing ability?*
*4.4 To what extent are exemplar essays chosen by means of the selected measures interpretable as a developmental scale of writing ability?*

### 1.2.3 Outline

In the present study, three approaches to the assessment of writing are evaluated, each in a separate chapter. Together, these chapters aim to answer the question of how to validly and reliably assess writing in primary education. Each chapter first offers an introductory section, followed by two or more sections in which the assessment approach under consideration is evaluated. The last section of each chapter discusses the results of the particular approach and offers a conclusion on its usability.

In Chapter 2, the use of rating scales with anchor essays is evaluated. Section 2.1 discusses the rationale for the use of this specific rating procedure. The process of constructing a rating scale with anchors is presented in Section 2.2. In Section 2.3, this assessment method is evaluated, and Section 2.4 discusses the results and their implications.

Chapter 3 concerns the use of a revision test as part of a writing assessment. In Section 3.1, the motive for the use of revision tests is explained, and Section 3.2 gives an overview of the construct of text revision and the different assessment methods. The use of a multiple-choice revision test is evaluated in Section 3.3, followed by an evaluation of a

constructed-response version in Section 3.4. In Section 3.5, the results of the use of revision tests are discussed.

Chapter 4 explores the use of automated evaluation of linguistic features. First, Section 4.1 offers an introduction on automated essay scoring and the validity of this method. In Section 4.2, the relation between text complexity and writing ability is explored, and appropriate measures are selected. Section 4.3 offers a qualitative study on these measures. Section 4.4 discusses the feasibility of automated essay evaluation within primary education.

Lastly, Chapter 5 provides an overall discussion of the results presented in previous chapters, as well as recommendations on the assessment of writing ability within a large-scale assessment in primary education.

# 2 Assessing text quality: The construction and evaluation of an anchored analytical assessment



## 2.1 Introduction

A productive ability, such as writing, can be assessed only by means of evaluating a candidate's performance. Writing ability is usually assessed through written products demonstrating the candidate's performance in a writing task. As illustrated in several studies over time (Breland, Camp, Jones, Morris & Rock, 1987; Godshalk, Swineford & Coffman, 1966; Cushing Weigle, 2002; Knoch, 2011), both the reliability and the validity of writing assessments can be questioned. The fact that raters disagree on the quality of a text, for instance, affects the reliability of a writing assessment, while the question of to what extent a rating procedure truly represents the construct of writing ability is a typical validity issue. Cushing Weigle (2002) and Shaw and Weir (2007) summarised a number of studies addressing these issues.

The present study focuses on improving the rating procedure within an assessment of writing. Although the influence of rating procedures on the scores assigned by raters is acknowledged (e.g. Schoonen, 2005), little research has been done on the effects of different marking methods (Barkaoui, 2011) and the results include both arguments for and against different rating methods (Cushing Weigle, 2002; Rijlaarsdam et al., 2012). This study aims at improving the reliability and validity of a large scale writing assessment in primary education by means of incorporating anchor essays into an analytical rating procedure.

### 2.1.1 Assessing writing in a large-scale assessment

In the 1980s, a national assessment program (PPON: *Periodieke Peiling Onderwijsniveau,* Dutch National Assessment in Education) was developed to evaluate the educational level in primary education in the Netherlands. Within this national assessment, a wide variety of topics is addressed, including writing. Several different writing tasks are undertaken by pupils in grade 3 and grade 6, in order to collect a large sample of writing products. These

writing samples are evaluated by one or more raters in order to determine the level of writing ability of the population. In order to be able to report on different aspects of writing, an analytical rating procedure is applied where each rater answers a list of evaluative questions per essay, implicitly assessing the aspects content, structure, and communication (Krom, Van de Gein, Van der Hoeven, Van der Schoot, Verhelst, Veldhuijzen & Hemker, 2004). One of the objectives of the specific type of analytical evaluation used in PPON is to alleviate the task of the raters by having them answer several straightforward yes/no questions on the essay, instead of explicitly evaluating certain aspects (cf. Brindley, 2001). Consequently, the raters only have to indicate whether specific features of the text are present (i.e. scoring), while the actual assigning of values to these features (i.e. grading) is done within the data analyses.

Because of the aforementioned simplification and objectification of the rating task, it was expected that this method would provide high agreement between raters (cf. Schoonen, Vergeer & Eiting, 1997). While this was indeed the case for the sub-aspects 'given content' and 'formal structure' (> .90), the analyses showed that the inter-rater agreement was relatively low (around .70) for several other aspects: 'generated content', 'content organisation', 'focus on public' and 'focus on writing goal' (Krom et al., 2004).

Apart from the 'generated content', these aspects dealt with more or less overall text features that concerned the essay as a whole, instead of specific parts of the text. A possible explanation for the disagreement between raters on these aspects would be that the raters do not get the opportunity to express their general impressions of the essay by simply answering the list of yes/no questions. It could therefore be the case that the raters 'misuse' their answers to the analytical questions to get their overall verdict of the essay across nonetheless, adding a source of disagreement to the evaluation and lowering the reliability of the assessment. Similar effects have been argued by Knoch, Read and Von Randow (2007) and Meuffels (1994).

In addition to the aforementioned reliability issue, the lack of overall judgment gives rise to questions of the validity of the assessment. Firstly, the raters may pay less attention to the written text as a whole by simply letting the questions lead them through the essay, causing text features to be isolated from the context (Perkins, 1983). Secondly, it is questionable whether the construct of writing ability can be 'captured' by a finite set of questions, meaning there is no guarantee that all relevant questions are being asked and, subsequently, that all relevant features of writing ability are assessed. Apart from the aforementioned possibility of raters compensating certain lacunae in the set of questions (by adjusting their analytical scores to match their overall impression of the essay; Knoch et al., 2007; Meuffels, 1994), findings from generalisability studies indicate that holistic scores better represent the complete construct of writing. That is, the generalisability of analytical writing scores across tasks and across raters is relatively low when compared to the holistic scores (Schoonen, 2005; Bergh, 2012). To obtain a satisfactory degree of generalisation, less raters and less writing tasks are needed when using holistic scores versus analytical scores.

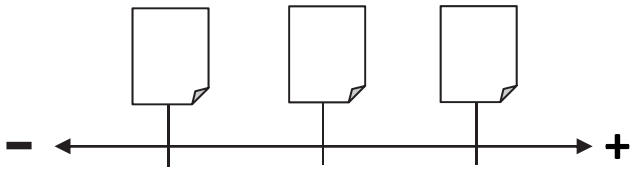



*Figure 1.* Schematic representation of a rating scale with anchor essays.

Given the above, the inter-rater agreement and validity of writing scores produced by an analytical rating procedure are expected to benefit from the addition of a holistic element, in which the essay as a whole is taken into consideration by the raters. This can be realized by providing raters with exemplar essays (so-called 'anchor essays' or 'benchmarks') which they can use to compare the quality of the writing products. The addition of the anchor essays as fixed reference points is expected to increase the agreement between raters. One way of incorporating these reference points into the rating procedure is to construct a scale representing writing ability, illustrated with several examples of writing products. These anchor essays are taken from a sample of essays and each represents a specific score point on that scale, ranging from a poor performance on one end of the scale to an excellent performance on the other end of the scale. The raters evaluate the overall quality of an essay for each aspect by placing the essay on this ability scale (cf. Figure 1).

Texts (1) and (2) illustrate the development in of writing ability in primary education. The texts are based on a writing task in which pupils are asked to argue that (even though they did not manage to collect enough coupons) they are entitled to receive a phone, by explaining that no more coupons were available. Text (1) is written by a pupil from grade 3 in primary education (age 8/9) and is flawed in several ways. Aside from some errors in spelling, the argument given in the letter is flawed and the voice in which the letter is written is not appropriate for the intended audience. Given these characteristics, Text (1) will score below average on language use, structure and content, respectively. Text (2) is written by a pupil from grade 6 in primary education (age 11/12), and demonstrates a higher level of writing ability, by giving a coherent argument and using a tone that is appropriate for an audience that consists of unfamiliar adults. Within the rating process, exemplar essays like (1) and (2) can be deployed as illustrations of the different ability levels within a population, hence providing a fixed reference point for raters.

(1) [grade 3, translated from Dutch]

> Hello Yummy firm I only have 8 coupons so I add 2 more wrappers because you cannot collect any more coupons because they are soldout so I cannot colect any more greetings ****

[original text]

> Hallo firma Smikkel ik heb maar 8 punten dus ik doe er nog 2 wikkels bij want je kunt geen punten meer sparen want ze zijn uit verkocht dus ik kan niet meer spaaren groetjes ****

(2) [grade 6, translated from Dutch]

> Dear Yummy,
>
> I recently saw the campaign to win a free phone, and so I thought it would be nice to join! To my surprise the coupons are nowhere to be found even though is not yet april thirty. I have put 8 coupons in the envelop and two complete wrappers without points, but actually I do expect the phone, because I cannot help it that they are gone.
>
> Kind regards,
> ****** *** ** ***

[original text]

> Beste Smikkel,
>
> Ik zag laatst de actie dat je een gratis telefoon kunt winnen, dat leek mij dus wel leuk om hieraan mee te doen! Tot mijn verbazing zijn er nergens meer spaarpunten te vinden ondanks het nog geen dertig April is.
> Ik heb in de envelop acht spaarpunten gedaan met twee hele wikkels zonder punten, maar eigenlijk verwacht ik nu wel die telefoon, want ik kan er niets aan doen dat er geen punten meer zijn.
>
> Met vriendelijke groet,
> ****** *** ** ***

### 2.1.2 Text quality and text structure

In addition to the aforementioned benefits, the proposed addition of anchor essays will arguably add to the validity of the assessment by allowing for a more righteous evaluation of text structure. This hypothesis is supported by the intuitive notion that text structure is key to text quality: in order to be considered well written, a text needs to be well structured. Without connectedness (or coherence), a text is merely a sequence of utterances which will have little meaning to the reader (Sanders & Spooren, 2007; Zwaan & Rapp, 2006). In other words, structure can be regarded as a 'constituting text principle' (Sanders & Schilperoord, 2006). Consequently, essays written by skilled writers are expected to be better organised than those written by less skilled writers, which makes text structure an essential aspect when assessing writing ability.

Another indication for the importance of text structure when evaluating writing is found in the theories on the cognitive processes of writing. In the 1980's, the focus of writing research shifted from the writing product to cognitive activities related to the complex activity of writing. Several models representing the writing process have been proposed over

the years, the most well-known being constructed by Flower and Hayes (1981), Bereiter and Scardamalia (1987), Hayes (1996) and Kellogg (1996). While each model has its own focus, all of the cognitive components can be grouped into the main activities of planning, formulating and revising (Van Weijen, 2008).

From the perspective of the writing process, the structure and content of a text reflect a specific part of the process, namely, the phase in which writing plans are made, ideas are generated, and information is structured (Van der Pool, 1995). In writing products, this phase is reflected in the text level, while the process of translating the ideas and plans into sentences is represented by the paragraphs and word level (macro, meso and micro-levels, respectively). Furthermore, a correspondence found between text structure and the developmental aspects of writing suggests that cognitive processes are reflected in the hierarchical structure of the text produced (Sanders & Schilperoord, 2006; Sanders et al., 1996; Sanders & Van Wijk, 1996ab).

Given the above, the content and overall structure should be adequately represented in a model for text evaluation, in order to assess a crucial part of the writing process (planning) and a crucial feature of texts (structure). Both of these text features are also addressed in the revision component, in which all aspects of the text produced thus far are evaluated and edited where necessary. Measuring this component of writing, however, requires a different method of assessment, and will therefore not be incorporated in this study.

Traditionally, linguists consider overt linguistic elements to depict connectedness between text elements, referred to as *cohesion* (cf. Halliday & Hassan, 1976). Following this view, the evaluation of text structure would be sufficiently covered by quantifying the use of these cohesive elements. The presence of explicit structural elements, however, is not a prerequisite for connectedness. Instead, more recent linguistic theories consider the connectedness of discourse to be a characteristic of the mental representation of the text, rather than of the text itself (Linterman-Rygh, 1985; Sanders & Pander Maat, 2006). This notion of connectedness is referred to as *coherence*. Language users themselves establish coherence by actively relating different information units in the text, during which they may or may not be assisted by cohesive elements.

The relationship between coherence and cohesion is illustrated by Texts (1) and (2). The text quality of these essays differs in several aspects, including text structure. Both essays contain cohesive elements that are relevant for the assignment, including the connectives *dus* (so), *want* (because), *ondanks* (even though) and *maar* (but). Despite the use of several cohesive features, Text (1) lacks coherence, which is why the argument presented is difficult for a reader to follow, and the letter is unlikely to reach its goal. Text (2), on the other hand, does not show as many cohesive elements, but rather, provides (implicit) textual relationships which help the reader to establish coherence.

Since simply counting cohesive devices does not suffice as an evaluation of text structure, a more sophisticated method is needed in which the coherence of a given text is

evaluated. To assist in this process, anchor essays will arguably prove to be helpful as a common reference point for raters. A valid evaluation of writing, in other words, should encourage raters to evaluate the writing product as a whole. In this study, providing raters with a rating scale with anchor essays, and asking them for an overall judgment per aspect of writing, is thus predicted to enable a valid assessment of writing ability.

### 2.1.3 Beneficial effects of rating scales with anchor essays

Van den Bergh and Rijlaarsdam (1986) developed a method to construct a rating scale with exemplar essays as mentioned above. The authors described all the steps needed to create rating scales for different aspects of writing. According to Van den Bergh and Rijlaarsdam, using a rating scale with anchor essays has two main advantages over the use of an analytical rating procedure. First of all, the exemplars on the rating scale serve as reference points, supporting the raters in their rating task and reducing instability in their rating. Moreover, using a fixed standard allows scores to be compared between pupils and classes or scores to be monitored over time. In the context of a national assessment, anchor essays can be particularly useful to illustrate different levels of achievement. In fact, anchor essays were used in earlier cycles of the national assessment for writing (Zwarts, 1990; Sijtstra, 1997), but were eventually replaced in the next cycle due to their complicated scoring instructions.

Relatively few studies have been conducted to examine the beneficial effects of using a rating scale with anchor essays when evaluating essays. The results published thus far seem to indicate a modest improvement in inter-rater agreement for methods using exemplar essays as anchors when assessing writing, compared to methods without exemplars. Rijlaarsdam (1986) used anchored rating scales to evaluate essays in a study on the effect of feedback given by peers vs. teachers and found a reliability of .80 between raters, which is considered to be reasonably high when assessing writing. Melse and Kuhlemeier (2000) compared two methods of rating (using an analytical rating scheme and using an analytical rating scheme supplemented with anchor essays) and reported an increase in reliability from .70 to .76 when using anchors. The practicality of rating scales with anchor essays is endorsed in several studies, including Blok and Hoeksema (1984), Schoonen and De Glopper (1992), Schwartz and Collins (1995), Van Gelderen, Oostdam and Van Schooten (2011), De Milliano (2012) and Sluijter, Verhelst and Hermans (1999). The latter resulted in the use of anchors to evaluate student drawings within the Dutch national assessment of visual arts (Hermans, Van der Schoot, Sluijter & Verhelst, 2001).

Furthermore, Schoonen (2005) and Van den Bergh, De Maeyer, Van Weijen and Tillema (2012) show that analytical scores are less generalisable than holistic scores as a measure of writing ability. Schoonen suggests that holistic scores include a general impression of writing ability and are less task-dependent. Van den Bergh et al. argue that while analytical scores are reliable, they are text and topic-dependent and, therefore, less indicative of general writing ability. These findings imply that although analytical scores are

appropriate to assess certain parts of writing ability in a reliable manner, holistic scores are needed in order to cover all aspects of writing ability.

Given the abovementioned characteristics of analytical and holistic scoring, this study presents a procedure in which both methods are integrated. In the so-called Anchored Analytical Assessment (AAA) procedure, a rating scale with anchor essays is added to an analytical procedure. The beneficial effects of this newly developed procedure are evaluated in this study. First, the five phases of constructing anchored rating scales are discussed (Section 2.2). In Section 2.3, the validity and reliability of the AAA procedure are evaluated, leading to a conclusion on its usability within the Dutch national assessment.

## 2.2 The construction of a rating scale with anchor essays

### 2.2.1 Introduction

To construct rating scales with anchor essays for the evaluation of essays written by primary school pupils, the procedure described by Van den Bergh and Rijlaarsdam (1986) was adopted for this study. The five consecutive phases of building the rating scales are: (1) specifying the features of the rating scale, (2) collecting the essays, (3) rating the essays, (4) analysing the rating data, and (5) constructing the scale. The result of this construction process is a rating scale with three exemplar essays for each assessed aspect of writing. Each of these anchor essays represents a specific score point on the rating scale, serving as a fixed reference point for raters. In this section, the different phases and steps in the construction of nine different rating scales with anchor essays are described.

### 2.2.2 Constructing a rating scale with anchor essays

*Pre-Phase – Specifying the features of the rating scale*

Before constructing the actual writing scale, decisions had to be made regarding the design of the scale. Cushing Weigle (2002) lists five topics that should be addressed when constructing a rating scale: 1) type of rating scale, 2) users of the scale, 3) aspects to be assessed, 4) type and number of descriptors, and 5) the reporting of scores. The features of the rating scale were specified according to these five topics.

Given the context of a national assessment, the rating scale must provide detailed information on the level of writing ability in the Netherlands. Therefore, an analytical scale for three separate aspects of writing ability is desired (1). The scale will be used by trained raters employed by Cito (2). Three previously defined aspects of writing were used, divided into several sub-aspects (cf. Table 4) (3). The descriptors will be actual essays, plus annotations listing the specific features of these essays. Per the rating scale, 3 descriptors are used, creating 7 scoring points (Figure 2) (4). Within the national assessment, no individual scores are reported. Finally, on the basis of all student responses, the overall scores per aspect will be converted into ability scores in order to evaluate the writing skills of Dutch primary school pupils (5).

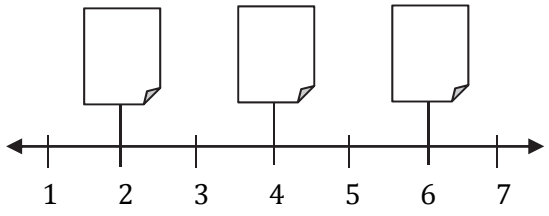



*Figure 2.* Scoring levels per rating scale.

*Phase 1 – Collecting the essays*

Step 1.1 – Formulating the assignment

Three different essay tasks were selected from the pool of tasks in the Dutch national assessment (Table 1). All tasks in the national assessment are classified by communicative goals (cf. Renkema, 1993), representing skills such as telling a story (narrative), making an appeal (directive) and providing arguments to convince the reader (argumentative). Also, a diverse collection of text genres is represented within the assessment. The three selected tasks represent a broad scope of text goals and text genres.

Table 1. *Essay Tasks*

| Task | Description | Communicative goal | Text genre |
|---|---|---|---|
| **A Tigors & Giraks** | Finishing an adventure story about two tribes | Narrative | Story |
| **B Pookie** | Writing a note describing a lost cat and requesting help | Descriptive, directive | Leaflet |
| **C Yummie** | Writing a letter to convince a company to accept an incomplete stamp card | Argumentative | Letter |

Step 1.2 – Piloting the assignment

All of the tasks were part of a previous cycle of the Dutch national assessment (Krom et al., 2004), and have been shown to elicit relevant student responses. Therefore, no additional pilot testing was arranged for this study.

Step 1.3 – Collecting the essays

A total of 270 Dutch primary schools were invited to participate in the study. To be able to use the test results as a comparative measure when evaluating the rating scale, only schools using Cito's Entrance Test were invited. Five schools, representing different regions, school sizes and denominations, volunteered to participate in the study. A total of 584 pupils participated, aged 8 to 12 (Table 2).

Table 2. *Collection of Essays*

| | Grade 3 Age 8/9 | Grade 4 Age 9/10 | Grade 5 Age 10/11 | Grade 6 Age 11/12 | Tasks per pupil |
|---|---|---|---|---|---|
| School 1 | 35 | 48 | 26 | 26 | 3 |
| School 2 | 25 | 25 | 25 | 25 | 3 |
| School 3 | 41 | 31 | 33 | 38 | 2 |
| School 4 | 48 | 34 | 44 | 24 | 2 |
| School 5 | 23 | 17 | 36 | 16 | 2 |
| **Essays per task** | **135** | **128** | **126** | **103** | |

In two schools, the participating pupils wrote essays for all essay assignments, i.e., three essays in total. In the other schools, the pupils wrote a set of two essays. To avoid sequence effects, one out of six possible combinations of tasks was assigned randomly (Table 3). In total, 1476 essays were collected. All essays were digitalised (i.e., typed over, maintaining layout, typos and punctuation) in order to facilitate reproduction and distribution. Moreover, the quality of handwriting can influence the assessment of text quality (De Glopper, 1988), and presenting the essays in typescript eliminates this unwanted effect.

Table 3. *Combinations of Essay Assignments*

| | **3 tasks per pupil** | **2 tasks per pupil** |
|---|---|---|
| 1 | A-B-C | A-B |
| 2 | A-C-B | A-C |
| 3 | B-A-C | B-A |
| 4 | B-C-A | B-C |
| 5 | C-A-B | C-A |
| 6 | C-B-A | C-B |

*Phase 2 – Rating the essays*

Step 2.1 – Formulating rating criteria

As in the previous cycles of the national assessment, three aspects of writing were rated (Table 4). The aspect 'Content' relates to the extent to which the content of the writing piece suits the specific task requirements, such as the content elements present in the text, focus on writing goal and focus on public. For 'Structure', the essays were evaluated on the way the content is organised, focusing on composition, coherence relationships between sentences and phrases, formal structure and layout. Finally, for 'Correctness', the form of the content was assessed by evaluating syntax, spelling and punctuation, as well as style. For each of the aspects, a separate rating scale was eventually developed.

Table 4. *Aspects of Writing*

| **Aspect** | **Description** | **Sub-aspects** |
|---|---|---|
| **Content** | Task requirements | Essential content elements<br>Extra content elements<br>Focus on writing goal<br>Focus on public |
| **Structure** | Composition, layout | Composition<br>Relations between sentences and phrases<br>Formal structure<br>Layout |
| **Correctness** | Syntax, spelling, punctuation | Syntax<br>Spelling and punctuation<br>Style |

Step 2.2 – Setting a standard

In order to familiarise the raters with the rating criteria, all four raters first assigned scores to a small random sample (n = 4) of essays for each task and aspect. The scores were then compared and discussed among the raters and the rating criteria were clarified where necessary. Furthermore, to have a fixed reference point, all raters decided upon an 'average' essay for each task and for each aspect. Agreement on the average essay was achieved by a discussion of the quality of the four essays among the raters. This essay was assigned 100 points and accompanied the agreed upon rating criteria (see Appendix A for an example).

Step 2.3 – Assigning the essays to raters

Out of the collection of essays, a random sample of 40 essays was drawn per task and for each aspect to be rated. A complete design was used: all raters rated all 360 essays. Three rounds of rating were performed: one round for each aspect. To avoid sequence effects for the tasks, the six possible orders in which the different tasks were to be rated were assigned randomly (Table 5). For example, when assessing the aspect Content, rater 1 first assessed task A, then task C, and finally task B. Rater 2 was assigned a different order, as were rater 3 and 4.

Table 5. *The Rating Design: Sequence of Tasks per Rater*

| | **Rater 1** | **Rater 2** | **Rater 3** | **Rater 4** |
|---|---|---|---|---|
| | Task | Task | Task | Task |
| **Round 1 - Content** | A | B | C | A |
| | C | A | A | B |
| | B | C | B | C |
| **Round 2 - Structure** | B | A | B | B |
| | C | C | A | C |
| | A | B | C | A |
| **Round 3 - Correctness** | B | C | A | A |
| | C | B | B | C |
| | A | A | C | B |

Step 2.4 – Evaluating the essays

Raters were instructed to assign a score to each essay, using the average essay (100 points) as a reference point, and thus indicating the degree to which the rated essay was considered to be better or worse. No further instructions were given on the score distributions: the raters were free to use a range of scores.

*Phase 3 – Analysing the rating data*

Step 3.1 – Converting the scores to standard scores

In order to be able to compare all of the scores independently of the score distribution, the scores given by all of the raters were converted to standard scores (z-scores). Next, these were transformed to a scale with a mean of 100 and a standard deviation of 15.

Step 3.2 – Evaluating the agreement

Using the standardized scores, the reliability of the scores was evaluated by determining the inter-rater agreement (correlation) between all four raters (Table 6). The agreement was considered to be sufficient to use the scores from all of the raters as a reliable source to select anchor essays.

Table 6. *Reliability of Rater Scores*

| | **A Tigors & Giraks** | **B Pookie** | **C Yummie** |
|---|---|---|---|
| Content | .93 | .88 | .85 |
| Structure | .88 | .80 | .87 |
| Correctness | .89 | .75 | .80 |

*Phase 4 – Constructing the rating scale*

Step 4.1 Selecting the anchor essays

All of the essays were ordered based on the standard scores. Exemplar essays were chosen around the score points of 100 (mean), 85 (-1 sd) and 115 (+1 sd) (Figure 3). Around these points, the essays were selected on the basis of the agreement between the four different raters by looking at the standard deviation of the scores. That is, to be chosen as an anchor, an essay had to represent one of the three score points listed above, plus the scores given to this essay by the raters had to vary as little as possible. Where necessary, i.e., when more than one essay qualified, the final decision was made on the extent to which an expert rater who had not been part of the team of raters believed the essay to be a 'typical' performance. Otherwise, the anchors were chosen on the basis of their empirically defined value as an exemplar essay. Appendix A lists all scores for each task and aspect.

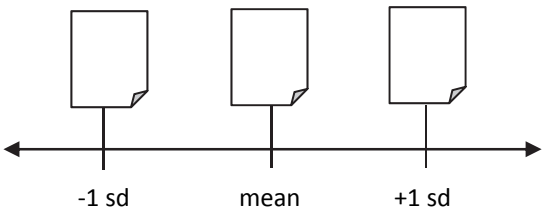



*Figure 3.* A rating scale with three exemplars.

Step 4.2 Annotating the rating scales

In addition to the three scales per task **(**one for each aspect of writing), short descriptions of the essay's characteristics (Figure 4; Appendix C) were added to the exemplars. This description was based on the criteria to be used by the raters when evaluating the essays.

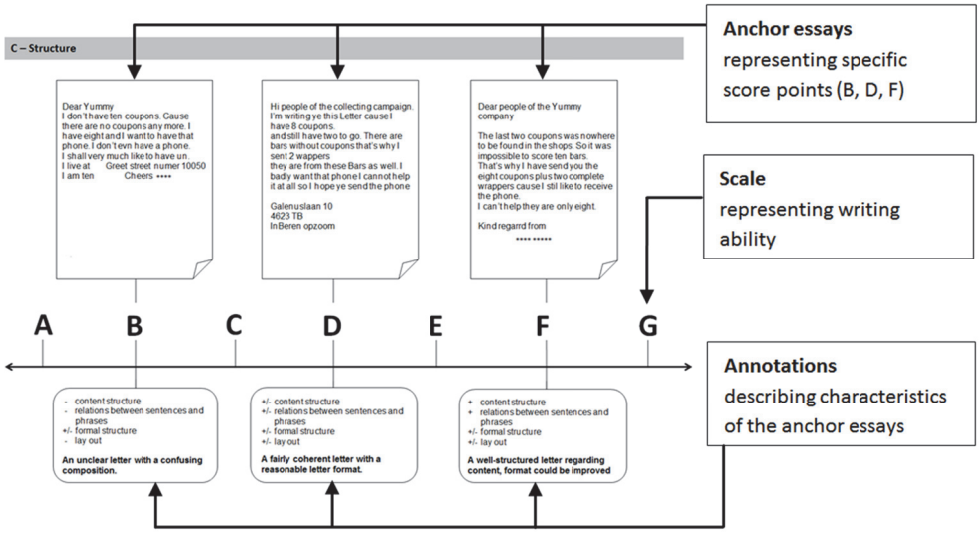



*Figure 4*. An annotated essay scale.

*Post-Phase – Evaluating the rating scale*

In the last phase, the rating scale was evaluated. In order to decide upon the usefulness of the newly developed rating scale with anchor essays, a design with two conditions was set up, comparing the adjusted scoring method (condition 2) to the existing method (condition 1). The test and item statistics, inter-rater reliability and relationship to external criteria were used to evaluate the quality of the new rating. Furthermore, the experiences of both groups of raters were collected. These analyses and their results are discussed in the following section.

## 2.3 The evaluation of an anchored analytical assessment of writing ability

### 2.3.1 Introduction

In this study, the procedure described by Van den Bergh and Rijlaarsdam (1986) was adopted to construct a rating scale with anchor essays for the evaluation of essays written by primary school pupils on the aspects of content, structure and correctness. The result of this construction process is a rating scale with three anchor essays for each of the three assessed aspects of writing. Each of these anchor essays represents a specific score point on the rating scale, serving as a fixed reference point for raters and thereby presumably increasing agreement between raters. In this study, the rating scale with anchors is combined with analytical questions, resulting in an 'anchored analytical assessment' (AAA) procedure. In this way, the assessment is expected to benefit from the advantages of both holistic and analytical scoring: attention to the 'wholeness' of writing performance (Roid, 1994) and an increased rater agreement on the one hand, and detailed analytical information on the other.

### 2.3.2 Research questions and hypotheses

The present study aims to explore the reliability and validity of the AAA procedure, resulting in a conclusion on its usability within a large-scale assessment. To do so, the inter-rater agreement, generalisability and construct validity of the AAA procedure is addressed by means of the following research questions:

1. *To what extent does the addition of anchor essays to an analytical evaluation provide higher inter-rater reliability when compared to a solely analytical procedure?*
2. *To what extent are writing scores obtained via the anchored analytical assessment generalisable across tasks and raters?*
3. *To what extent do writing scores obtained via the anchored analytical assessment provide evidence for construct validity?*

First of all, when assessing writing, the agreement between raters is a precondition for the reliability of the assessment. It is expected that having anchor essays as a reference will help raters to agree on the quality of the essays, resulting in higher inter-rater reliability when compared to a solely analytical evaluation. Second, writing scores should be generalisable across tasks and across raters, thus representing all possible tasks and raters, and indicating proper coverage of the construct of writing ability. The anchored analytical assessment is predicted to diminish the effect of unwanted sources of variance, such as the task or rater on the generated scores, therefore resulting in highly generalisable writing scores. Third, evidence of construct validity is needed to ensure that the targeted construct is addressed. With respect to this, the average essay scores per grade serve as measures of known-group validity, as writing ability is known to increase with grade. In addition, correlations between writing scores and other (external) measures of language ability are expected to be higher in comparison to correlations with scores on arithmetic, serving as measures of convergent and

divergent (discriminant) validity, respectively. Finally, the scores for each aspect of the writing construct should be highly correlated *across* tasks in order to support construct validity, while the scores for different aspects within tasks are expected to correlate less strongly, indicating that different aspects can indeed be identified within writing ability.

### 2.3.3 Method

Materials and participants

Three different essay tasks were selected from the pool of tasks in the Dutch national assessment (Table 7). Writing performances are known to vary across tasks (cf. Bergh et al., 2012); therefore, to optimise the representation of the writing construct, the three assignments were selected to cover a broad scope of communicative goals and text genres.

Table 7. *Essay Tasks*

| Task | Description | Communicative goal | Text genre |
|---|---|---|---|
| **A Tigors & Giraks** | Finishing an adventure story about two tribes | Narrative | Story |
| **B Pookie** | Writing a note describing a lost cat and requesting help | Descriptive/ Directive | Leaflet |
| **C Yummie** | Writing a letter to convince a company to accept an incomplete stamp card | Argumentative/ Persuasive | Letter |

All essays were digitalized (i.e., retyped, maintaining layout, typos and punctuation) to facilitate reproduction and distribution. Moreover, handwriting quality can influence the assessment of other aspects of text quality (De Glopper, 1985), and presenting the essays in typescript eliminates this unwanted effect. As in the previous cycles of the national assessment, three main aspects of writing were to be rated, as shown in Table 8. Two sub-aspects were added for Structure and the sub-aspects for Correctness were re-ordered into three separate categories.

Table 8. *Aspects of Writing*

| Aspect | Description | Sub-aspects old | Sub-aspects new |
|---|---|---|---|
| **Contents** | Task requirements | Essential content elements | = |
| | | Additional content elements | = |
| | | Focus on writing goal* | = |
| | | Focus on public* | = |
| **Structure** | Composition, layout | Content organisation* | = |
| | | - | Relationships between sentences and phrases |
| | | Formal structure | = |
| | | - | Layout |
| **Correctness** | Syntax, spelling, punctuation | Language use | - |
| | | - | Syntax |
| | | *S*pelling | Spelling and punctuation |
| | | - | Style |

* Aspects generating a relatively low inter-rater agreement (Krom et al., 2004)

The procedure described by van den Bergh and Rijlaarsdam (1986) was adopted to compose a rating scale with anchor essays for each aspect per task, the result being nine rating scales in total, each with three anchor essays representing specific ability levels (Figure 5).

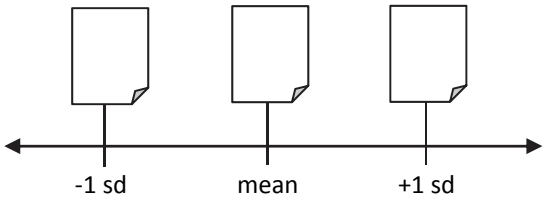



*Figure 5.* A rating scale with three exemplars.

To select the anchor essays, four expert raters first agreed upon the average essay and then evaluated a sample of essays (cf. Section 2.2). Their scores were converted into standard scores (z-scores). Three anchor essays (cf. Figure 5) were then selected based on their empirically defined values as exemplar essays, with high agreement among the four different raters. See Appendix A for an example of a rating scale with anchors for the aspect Structure.

Five Dutch primary schools representing different regions, school sizes and denominations volunteered to participate in this study. A total of 620 pupils, aged 8 to 12, participated. In two schools, the participating pupils wrote essays for all essay assignments (i.e., three essays in total). In the other schools, the pupils wrote a set of two essays. To avoid sequence effects, one out of six possible combinations of tasks was assigned randomly. In total, 1,475 essays were collected (Table 9).

Table 9. *Collection of Essays*

| | **Grade 3 Age 8/9** | **Grade 4 Age 9/10** | **Grade 5 Age 10/11** | **Grade 6 Age 11/12** | **N pupils total** | **Tasks per pupil** | **N essays total** |
|---|---|---|---|---|---|---|---|
| School 1 | 35 | 48 | 26 | 26 | 135 | 3 | **405** |
| School 2 | 25 | 25 | 25 | 25 | 100 | 3 | **300** |
| School 3 | 41 | 31 | 33 | 38 | 143 | 2 | **286** |
| School 4 | 48 | 34 | 44 | 24 | 150 | 2 | **300** |
| School 5 | 23 | 17 | 36 | 16 | 92 | 2 | **184** |
| **N pupils total** | **172** | **155** | **164** | **129** | | | **1475** |

Data analyses

In order to decide upon the quality of the newly developed rating scale with anchor essays, a design with two conditions was set up, comparing the adjusted scoring method (analytical questions with anchor essays, condition 2) to the existing method (solely analytical questions, condition 1). All raters were randomly assigned to one of the two conditions. In each of the two conditions, all of the essays were compiled in batches of 50. Each rater scored one batch per task and each batch was rated by a minimum of two raters. An overview of the rating design is given in Appendix B. To avoid sequence effects, the order in which to score the batches (1st, 2nd, 3rd) was assigned randomly for each rater.

In condition 1, the raters used only the existing list of analytical questions to evaluate their essays (i.e. one list per task). In condition 2, the analytical questions were ordered per aspect: the raters were given one list of analytical questions accompanied by a rating scale with anchor essays per aspect (i.e. three in total per task). In the final question per aspect, they were asked to provide an overall score (A to G) for this particular aspect by placing the essay on the scale (cf. Figure 6).

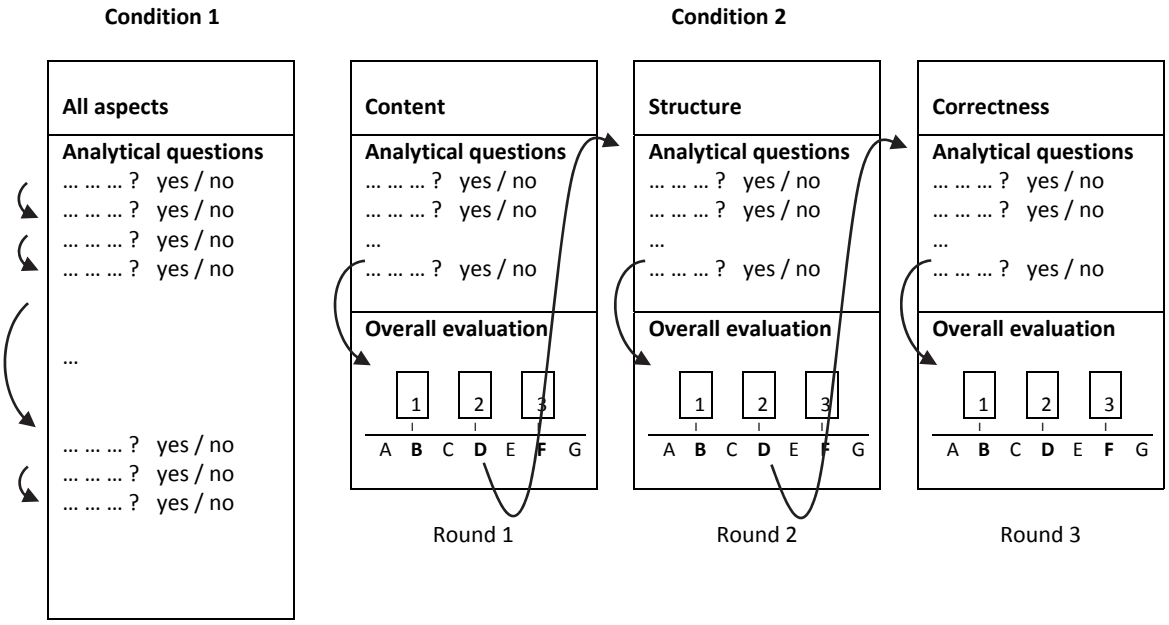



*Figure 6.* Schematic representation of the rating procedure for both test conditions.

The existing rating procedure (Condition 1) consists of a collection of questions about the properties of the essay to be evaluated; mostly dichotomous questions (yes / no) and some trichotomous (yes / yes, but ... (incomplete) / no). However, these trichotomous questions were eventually collapsed into dichotomous questions during the analyses, because only two out of the three categories proved informative when discriminating between the achievement levels. Therefore, only dichotomous questions were used in the adapted

procedure. All trichotomous questions in condition 1 were converted into dichotomous questions by collapsing two of the three answering options into one. For example, the answers “yes” and “yes, but …” would fall into one category, while the answer “no” would be the second category.

List-wise deletion was used to deal with the missing values in the dataset. In other words, when responses from one rater were missing for a certain essay (other than missing by design), this essay was left out of the analyses. Furthermore, two newly developed analytical items were deleted from condition 2 because of very poor item statistics.

The agreement between the raters was determined with the use of Gower’s coefficient; a measure specifically designed to deal with data lacking variance. A paired t-test was applied to test whether the differences found between the two conditions were significant. The program Tiaplus® (Heuvelmans, 2011) was used to compute various test and item statistics, as well as correlations with previous test results as external criteria.

To explore the generalisability of writing scores for the two conditions, the variance in the observed score was decomposed into components representing the different sources of variance in the research design, namely: person, task, rater, method, their mutual interactions and error (i.e. the interaction between all components). Estimates for each source of variance were created by means of the restricted maximum likelihood (REML) using SPSS (version 21).

### 2.3.4 Results

Inter-rater reliability

Table 10 presents the inter-rater agreement per aspect and per task. Each essay was scored in two conditions: condition 1 represents the existing analytical rating procedure, and condition 2 represents the adjusted version of the original procedure, consisting of an adjusted version of the analytical scale plus the additional rating scale with anchor essays.

Table 10. *Inter-Rater Agreement per Task and Aspect for All Grades (3 to 6)*

| | | Condition 1 (old) | Condition 2 (new) |
|---|---|---|---|
| CONTENT | | | |
| **A – Tigors & Giraks** | Average | **.84** | **.78** |
| Narrative | Std. Dev. | .03 | .05 |
| | Min. | .80 | .68 |
| | Max. | .88 | .85 |
| **B – Pookie** | Average | **.86** | **.86** |
| Directive | Std. Dev. | .07 | .07 |
| | Min. | .73 | .70 |
| | Max. | .90 | .92 |
| **C – Yummie** | Average | **.87** | **.87** |
| Persuasive | Std. Dev. | .01 | .02 |
| | Min. | .85 | .83 |
| | Max. | .89 | .89 |
| | **Average** | **.85** | **.84** |
| STRUCTURE | | | |
| **A – Tigors & Giraks** | Average | **.78** | **.80** |
| Narrative | Std. Dev. | .06 | .05 |
| | Min. | .67 | .74 |
| | Max. | .86 | .87 |
| **B – Pookie** | Average | **.78** | **.79** |
| Directive | Std. Dev. | .06 | .08 |
| | Min. | .68 | .64 |
| | Max. | .83 | .89 |
| **C – Yummie** | Average | **.72** | **.84** |
| Persuasive | Std. Dev. | .04 | .07 |
| | Min. | .68 | .69 |
| | Max. | .78 | .90 |
| | **Average** | **.76** | **.81*** |
| CORRECTNESS | | | |
| **A – Tigors & Giraks** | Average | **.75** | **.79** |
| Narrative | Std. Dev. | .05 | .06 |
| | Min. | .67 | .67 |
| | Max. | .80 | .87 |
| **B – Pookie** | Average | **.80** | **.75** |
| Directive | Std. Dev. | .03 | .09 |
| | Min. | .76 | .58 |
| | Max. | .84 | .88 |
| **C – Yummie** | Average | **.74** | **.77** |
| Persuasive | Std. Dev. | .11 | .07 |
| | Min. | .54 | .64 |
| | Max. | .88 | .84 |
| | **Average** | **.76** | **.77** |

*: significant (p = .008)

Generalisability

To explore the sources of variance that affect the writing scores for both conditions, the observed score variance was decomposed into the variance components of person, task, rater, their mutual (pairwise) interactions and error (i.e. the interaction between all components and measurement error). Since the AAA procedure is composed of two scoring methods (i.e. analytical questions and anchor items), 'method' was added as a variance component. Table 11 and Table 12 show the percentages of the variance in scores accounted for by each component per condition.

Table 11. *Variance Components of Observed Score for Condition 1 (old)*

| **Aspect** | **Content** | **Structure** | **Correctness** |
|---|---|---|---|
| **Variance component** | **%** | **%** | **%** |
| Person | 8.44 | 7.73 | 0.00 |
| Task | 59.67 | 38.84 | 1.63 |
| Rater | 0.00 | 0.40 | 5.50 |
| Person*Task | 17.21 | 27.64 | 9.84 |
| Person*Rater | 0.61 | 0.66 | 7.99 |
| Task*Rater | 0.41 | 0.72 | 1.63 |
| Random error | 13.67 | 24.01 | 73.41 |

Table 12. *Variance Components of Observed Score for Condition 2 (new)*

| **Aspect** | **Content** | **Structure** | **Correctness** |
|---|---|---|---|
| **Variance component** | **%** | **%** | **%** |
| Person | 22.80 | 33.40 | 44.30 |
| Task | 0.90 | 0.20 | 0.00 |
| Rater | 0.10 | 0.00 | 0.00 |
| Person*Task | 21.80 | 13.30 | 6.90 |
| Person*Rater | 0.00 | 2.50 | 0.00 |
| Task*Rater | 0.40 | 0.40 | 0.50 |
| Person*Method | 4.90 | 4.70 | 3.10 |
| Task*Method | 1.70 | 4.20 | 0.10 |
| Rater*Method | 0.00 | 0.00 | 0.20 |
| Person*Rater*Method | 0.00 | 0.00 | 0.90 |
| Person*Task*Method | 12.20 | 6.40 | 2.00 |
| Task*Rater*Method | 0.60 | 0.30 | 0.00 |
| Person*Task*Rater | 15.20 | 18.40 | 31.40 |
| Random error | 19.20 | 16.30 | 10.60 |

The results show that within the existing procedure (Table 11), only a small proportion of variance in writing scores is accounted for by the difference in ability between writers, while the specific task assigned accounts for a large proportion of variance when assessing content and structure. For correctness, a person's score seems to be based on the interaction of all variance components (i.e. error) instead of the writer's ability. Within the AAA procedure (Table 12), the effect of task is negligible, and the difference in writing ability accounts for the largest proportion of the variance in scores.

In Appendix E, the results for both components of the AAA procedure (i.e. the analytical questions and the anchor items) are considered separately; both showing a similar pattern, resembling the results for the combined method. While the task is not of great influence on the writing scores, the scores do differ greatly among the raters. However, these results can only be regarded as indicative since both components of the AAA procedure were (deliberately) intertwined within the design (cf. Figure 6).

Based on the variance components in Table 12, the generalisability of the absolute writing scores obtained by the AAA procedure was estimated. A writing score that is a perfect representation of the performances, based on all possible tasks, raters, methods and their combinations, would be optimally generalisable to writing ability in general, and hence, receive a generalisability score of 1. In Figure 7, the estimated generalisability coefficient per aspect (content, structure, correctness) is visually represented for an increasing number of tasks (x-axes) and raters (lines).

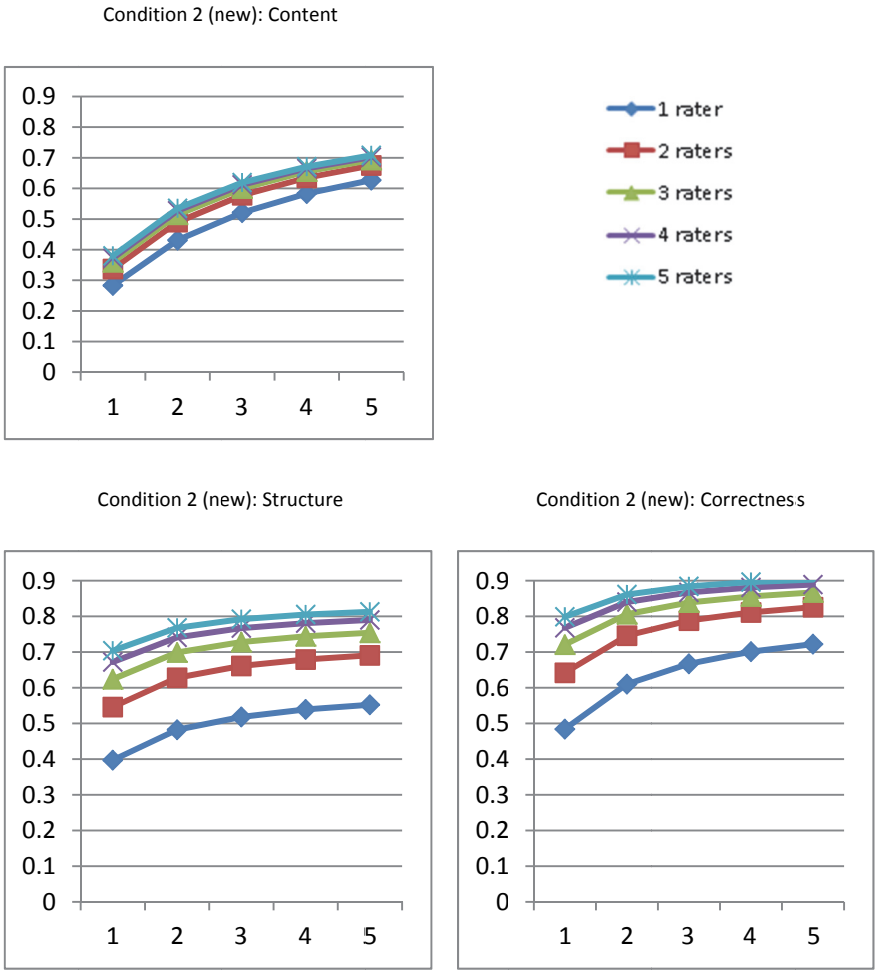



*Figure 7.* Estimated Generalisability Coefficient of Writing Scores.

The generalisability coefficients presented in Figure 7 indicate the degree to which the test scores obtained within the design in this study are generalisable across different tasks,

raters, methods and their interactions. As such, this coefficient can be interpreted as an equivalent of the reliability coefficient in classical test theory. Hence, a generalisability coefficient of .70 is considered to be 'good' for assessment at the group-level (cf. COTAN, 2010). Given this standard, the results presented in Figure 7 can be used to determine the number of tasks and raters needed to obtain a satisfactory degree of generalisation for the writing scores.

The results indicate that when assessing the aspect of Content, as many as 5 raters and 5 different tasks are needed to obtain writing scores that are generalisable (≤ .70) across raters and tasks. For Structure, the desired degree of generalisability can be realized by a smaller number of tasks (1 task and 5 raters) or a smaller number of raters (2 raters and 5 tasks). With respect to Correctness, as few as 1 task evaluated by 3 raters or 5 tasks evaluated by 1 rater will provide for a satisfactory degree of generalisability.

Construct validity

To further explore the validity of the AAA procedure, the p-values across grades and correlations across tasks and aspects were evaluated. Since writing ability increases with age, the performance of the writing tasks should improve across grades. Hence, the writing scores per grade can serve as a measure of 'known-group' validity. Figure 8 shows the proportions of points scored (p-values, averages for all tasks) per grade in both conditions. Separate tables showing all item characteristics per task and per grade for both conditions can be found in Appendix F and Appendix G.

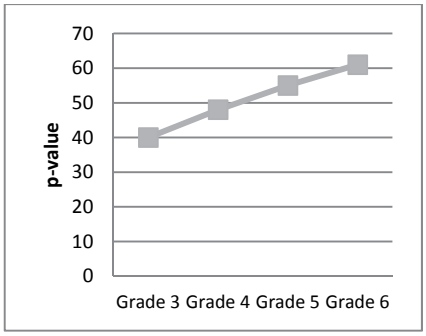



*Figure 8.* Average p-values for all tasks.

Construct validity of the AAA procedure was further explored by considering the correlations with both the related and unrelated constructs. Serving as measures of convergent and divergent validity, the correlations with related constructs (i.e. other language assessments) are expected to be higher when compared to correlations with unrelated constructs (i.e. arithmetic). Table 13 reports the correlations of the writing scores with the test results in different sections of the Cito Entrance Test as an external criterion of construct validity. Since no entrance test is available for grade 6, the results mentioned only

concern grades 3 to 5. Correlations between writing scores and the scores on three language subjects (i.e. text revision, vocabulary and reading) were compared to the correlations with arithmetic. The results show that the writing scores obtained via the AAA procedure show a significantly higher correlation with other language subjects when compared to the correlation with arithmetic.

Table 13. *Correlations With External Criteria per Task (grades 3 to 5)*

| Condition 2 (new) | N | Text revision | Vocabulary | Reading | Arithmetic |
|---|---|---|---|---|---|
| Task A | 273 | .59*** | .57** | .57** | .45 |
| Task B | 269 | .49*** | .42*** | .44** | .30 |
| Task C | 250 | .52*** | .52* | .48** | .39 |

***p ≤ 0.001; **p < 0.01; *p < 0.05 (given correlation vs. correlation with arithmetic)

Table 14 shows the correlations between the sum scores per aspect of writing. The correlations were corrected for attenuation, which caused some correlations to exceed 1.00, indicating an underestimation of reliability (cf. Lord & Novick, 1968). In these cases, the correlations were rounded-down to 1.00. The correlations between the different aspects within the same task (in *italics*) indicate the degree to which the different aspects of the writing performance can be identified *within* a performance on a specific task. The correlations between the aspects across different tasks (in **bold**) indicate the degree to which the scores on these aspects are related *across* performances on different tasks. In addition, Appendix H reports the correlations for the newly developed analytical items and anchor items separately. The results show that the correlations within the tasks are generally high (ranging from .61 to 1.00), while the correlations across the tasks are fairly low (ranging from .23 to .78).

Table 14. *Average Correlations[a] across All Grades, Condition 2 (new)*

| | | Task A | | | Task B | | | Task C | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Aspect** | 1 | 2 | 3 | 1 | 2 | 3 | 1 | 2 | 3 | Reliability[b] |
| **A** | **1 – Content** | *1.00* | | | | | | | | | 0.81 |
| | **2 – Structure** | *1.00* | *1.00* | | | | | | | | 0.81 |
| | **3 – Correctness** | *1.00* | *1.00* | *1.00* | | | | | | | 0.36 |
| **B** | **1 – Content** | **0.53** | 0.52 | 0.75 | *1.00* | | | | | | 0.77 |
| | **2 – Structure** | 0.60 | **0.73** | 0.84 | *0.89* | *1.00* | | | | | 0.71 |
| | **3 – Correctness** | 0.45 | 0.56 | **0.23** | *0.61* | *0.84* | *1.00* | | | | 0.84 |
| **C** | **1 – Content** | **0.49** | 0.55 | 0.43 | **0.44** | 0.37 | 0.30 | *1.00* | | | 0.88 |
| | **2 – Structure** | 0.66 | **0.74** | 0.93 | 0.54 | **0.63** | 0.44 | *0.96* | *1.00* | | 0.78 |
| | **3 – Correctness** | 0.45 | 0.64 | **0.78** | 0.38 | 0.51 | **0.31** | *0.64* | *0.91* | *1.00* | 0.69 |

[a] Corrected for attenuation
[b] Spearman Brown

## 2.4 Discussion and conclusion

In this study, the Anchored Analytical Assessment (AAA) procedure was presented: a newly developed method to assess writing ability, combining analytical and holistic scoring. Below, the results of the inter-rater reliability, validity and generalisability are discussed, leading to a conclusion on the usability of the AAA procedure.

Inter-rater reliability

The results of the inter-rater agreement given in Table 10 show that not all of the writing tasks and not all of the aspects of writing seem to benefit equally from the use of the AAA procedure. For the aspects Structure, the addition of anchor essays to the analytical rating scale appears to generally result in a higher inter-rater agreement, thus improving the reliability of the assessment. The assessment of content and correctness does not appear to benefit from the use of anchor essays. Thus, the present study shows that adding an overall judgment based on the comparison with exemplars improves the quality of the evaluation of Structure. Since the text structure is considered to be an important and useful aspect when evaluating writing (De Glopper, 1996; Sanders et al., 1996; Sanders & Schilperoord, 2006; Van der Pool, 1995), an improvement in the inter-rater agreement for the assessment of structure is regarded as a meaningful improvement within the assessment of writing.

The given results indicate that for the evaluation of text structure, in particular, having a complete essay as a reference for scoring is crucial, meaning that the analytical questions alone would not be sufficient to evaluate the structure. The evaluation of the aspects Content and Correctness, on the other hand, seem to be less dependent on an overall judgment. Apparently, the analytical questions enable the raters to express their complete judgment of these aspects of the essay, more or less objectively, and no room is needed for further interpretations.

A possible explanation for these differences in the effect on reliability could be the fact that the aspects Content and Correctness are largely built up from elements that can be tallied individually (e.g. number of required content elements present, number of grammatical errors). When assessing Structure however, rating is not just a matter of counting independent cohesive units, such as reference words and conjunctions, since text structure is not necessarily marked explicitly by cohesive devices. Rather, the connectedness within a specific text is composed of all explicit and implicit structural elements present in this text. Together, these elements reflect the *coherence* of the text. Consequently, the text as an entity is to be evaluated in order to validly assess text structure. In the present study, this prerequisite has been met by using a rating scale with anchor essays within the AAA procedure (seemingly resulting in improved reliability).

Generalisability

When assessing writing, the combination of a writer (person) and a given assignment (task) leads to a specific writing product. This product is then evaluated by one or more persons

(rater), following certain instructions (method), and resulting in a judgment of writing ability (score). Ideally, all variance in writing scores across writers is accounted for by the ability of the specific writers (person), regardless of the specific task and/or rater assigned. All other sources of variance (task, rater, method) can be considered unwanted effects, or measurement error. In other words, a good writer is expected to generally perform well, irrespective of variables such as task, rater or rating method. However, these variables are known to influence writing scores, and hence, affect the reliability and validity of a writing assessment. With the use of generalisability theory (Cronbach, Gleser, Nanda & Rajaratnam, 1972), the different sources of measurement error were estimated for both methods. Table 11 and Table 12 present the percentages of variance in scores accounted for by different components within the design.

The results show the large effect of task on the solely analytical procedure (Condition 1) when compared to the AAA procedure (Condition 2). Apparently, the writing scores in Condition 1 differ greatly per task, indicating a difference in 'difficulty' for the analytical items per task (i.e. average scores assigned by all raters differed between the three tasks). Furthermore, a large interaction between person and task was found for both conditions, indicating that the writing performance is highly influenced by the specific task assigned to an individual. This effect of task is partly anticipated, since, in this study, task and genre are intertwined. That is, every task represents a different genre of writing (cf. Table 7). Therefore, the variance in writing scores accounted for by the difference in ability is likely to be underestimated, as persons are known to perform differently across genres (cf. Bouwer, Béguin, Sanders & Van den Bergh, in press). This notion is supported by the fact that for both conditions, the influence of task is highest when assessing Content, which includes several items that are task specific, hence related to a specific genre in this design. The aspects Structure and Correctness are expected to represent features of writing for which performance is largely independent of task, which seems to be supported by the fact that the scores of these aspects are less affected by task.

To further explore the sources of variance for the AAA-procedure, the variance components for both the analytical questions and the anchor items were considered separately (Appendix E). However, since both methods were entwined in the actual rating procedure, these results can only be interpreted as an indication of the possible results when employing one of the two methods separately. Subject to this reservation, the results in Appendix E suggest a smaller influence of task and rater for the newly developed analytical questions, when compared to the existing rating procedure (cf. Table 11).

These results can be explained in two ways. Firstly, given the fact that both procedures were integrated (cf. Figure 6), it is likely that the analytical scores were (partly) influenced by the presence of a rating scale with anchor essays. That is, after several rounds of scoring, raters will probably have 'internalised' the exemplars, which apparently improves the writing scores by lowering the unwanted influence of the specific task assigned. Secondly, within the AAA-procedure, the three different aspects of writing were evaluated

explicitly and in different runs of the procedure, while in the existing procedure the aspects were evaluated implicitly and in the same run (cf. Figure 6). The reported results suggest a positive influence of explicitly indicating the different aspects to be evaluated by addressing them in different rounds. Possibly, the performances on different aspects interfere with the existing procedure, leading to interdependent scores (cf. halo-effect, Meuffels, 1994); hence, lowering the amount of variance accounted for by the writer.

Based on the estimated variance components (cf. Table 12), Figure 7 reports the estimated generalisability of the writing scores for different numbers of tasks and raters. These results show that the AAA procedure items generate scores that are generalisable (> .70) across tasks and raters when at least five tasks scored by five raters are administered while assessing Content. For the aspects Structure and Correctness, a combination of, respectively, 3 tasks and 3 raters and 2 tasks scored by 2 raters is needed to be able to generalise across tasks and raters.

Together with Table 11 and Table 12, the results of this generalisability analysis indicate that the use of anchored analytical assessment essays leads to scores that are more generalisable when compared to scores on solely analytical items. This difference is possibly caused by the fact that the AAA procedure incorporates anchor scores, which are based on an overall impression of writing performance, whereas the scores on the analytical items rely on specific, more local features of the essays. These features might not be equally indicative of general writing ability. Besides, analytical items force raters to judge only the features questioned by the specific items, whereas anchor items enable raters to capture more facets of writing ability than explicitly mentioned.

Construct validity

A comparison of the expected test results for the 'known groups' with actual results for pupils belonging to these groups yields evidence for the construct validity of the AAA procedure. In this case, different grades in school are considered to be the known groups: their ability in writing is known to grow each year. Comparing the average difficulty of the writing assessment for these groups indeed shows the expected increase in writing ability (cf. Figure 8), hence supporting the validity of the AAA assessment.

Furthermore, Table 13 reports the correlations between the essay scores and scores on measures of related constructs (i.e. language constructs) and one unrelated construct (i.e. arithmetic). These results are to be interpreted as measures of construct validity, since theoretically related constructs are assumed to be interrelated in reality, showing a higher correlation when compared to correlations with unrelated constructs (convergent and divergent validity, respectively, cf. Cook & Campbell, 1979). Measures of the addressed construct, writing in the present case, should therefore show a moderate correlation with measures of similar constructs (i.e. other language tests) and a significantly lower correlation with measures of different constructs (i.e. arithmetic). This pattern is found,

indicating that a language construct is indeed measured and thus adding evidence for the construct validity of the procedure.

In Table 14, correlations between the scores on different aspects and tasks are presented. These correlations can be interpreted as measures of convergent and divergent validity: correlations between the scores on the same aspect for different tasks are expected to be higher than those for scores on different aspects within the same task. However, the results for both conditions show that the correlations within a specific task and across aspects are typically high, while correlations within aspects and across tasks are relatively low. These results are in line with previous research reporting covariance of text quality traits (De Glopper, 1985; Van den Bergh, 1988; Godshalk et al., 1966; Lee, Gentile & Kantor, 2008; McNamara, 1990) and indicate that, based on these scores, both differentiation between different traits within writing ability and generalisation across different tasks are problematic. However, as Deane and Quinlan (2010) point out, the separate traits of writing do reflect the specific targets of writing instruction, justifying a differentiated report on writing ability.

Apart from the abovementioned evidence of convergent and divergent validity, the fact that the essay as an entity is explicitly taken into account in the AAA procedure adds to the validity of the scores derived from this assessment. Structure is a text feature that distinguishes a collection of utterances from a meaningful text. Moreover, the hierarchical structure of a text develops in line with the developmental aspects of writing (Sanders & Schilperoord, 2006; Van der Pool, 1995). The evaluation of structure should therefore be considered an important aspect within an assessment of writing ability. Since the coherence of a text is most naturally assessed when evaluating the text as a whole, the approach as applied in the AAA procedure arguably leads to a higher validity when compared to a solely analytical procedure. A similar conclusion is drawn by Bamberg (1984), who reports on the holistic evaluation of coherence per text, compared to the local evaluation of coherence per paragraph. When taking the whole text into account, larger differences between age groups where found. Furthermore, a high correlation between holistic coherence scores and overall essay quality was found, indicating that overall coherence is indeed an important feature to evaluate.

In addition, when reading a text, raters – knowingly or unknowingly – form a holistic impression of that text. When answering analytical questions only, raters may therefore feel their answers for the analytical questions do not match their overall impression, and might, consequently, adjust their scores to match their holistic impression (cf. Knoch et al., 2007; Meuffels, 1994). By combining both analytical questions and an overall comparison with anchor essays in the AAA procedure, all relevant text features are explicitly assessed, while enabling raters to express their overall impression of the text as well, hence, adding to the validity of the assessment.

Usability

The success of a rating method is highly dependent on the proper execution by its users, that is, the raters. Both groups of raters were asked to evaluate the rating procedure they used. Raters did not report a strong preference for the AAA procedure, although they did consider the anchor essays to be a useful reference point. The training session was evaluated as useful by almost all of the raters. They especially appreciated discussing the difficulties of the job with the other raters and finding out that all raters are, to some extent, insecure about their judgments. Certain raters found it difficult to fully adopt the new rating method and reported that they sometimes forgot to consider the anchor essays while rating. It could, therefore, be beneficial to give more attention to the training of raters to make sure that the method is fully and correctly adopted by all raters. Some raters also mentioned the fact that it takes time to incorporate the comparison to anchor essays within their rating practice.

Another aspect of usability to take into consideration is the amount of resources needed to execute the AAA procedure. In this study, the raters first answered a list of analytical questions per aspect, leading to the placement of the essay on the scale in the final question for each aspect. When considering the scores for the final questions separately, however, an inter-rater agreement of .82 on average, over all aspects, was found, indicating that these questions as stand-alone items provide reliable information on the ability of the test-taker. Furthermore, decomposition of the variance components (Appendix E) shows that the anchor items provide scores with a high level of generalisability. All in all, these results indicate that the anchor items are reliable and informative assessment units. These scores, however, cannot be interpreted reliably as of yet, because of the influence that the answering of the analytical questions will have on the judgment made in the final question. Despite this, these one-item assessments look promising and might well be developed into useful tools for classroom assessment because of their efficiency (cf. comparative judgment, Pollitt, 2004).

Finally, one of the raters reported that the AAA-procedure highlights the fact that discrepancies can exist between the quality 'computed' by answering the yes/no-questions on the one hand, and the quality observed when looking at the essay as a whole, on the other. Certain essays 'tick all the boxes' for a specific aspect, but still, something about the text indicates that the writer has not yet mastered the aspect. Instead of forcing raters to ignore the discrepancies, the use of a rating scale with anchor essays acknowledges this phenomenon, hence, allowing for a more reliable and valid evaluation of writing.

## Conclusion

This study presents the newly developed 'anchored analytical assessment' (AAA) procedure, which combines analytical questions with scores based on anchor essays for each of the aspects content, structure and correctness. The aim of this study was to explore whether the addition of anchor essays leads to a more reliable and valid assessment of writing. A comparison of the AAA procedure to a solely analytical assessment procedure has shown

that the addition of anchor essays provides for significantly higher agreement between raters when assessing text structure. Furthermore, the analyses of convergent and divergent validity, as well as the results of a multiple regression analysis and a generalisability study, indicate that the use of anchor essays adds to the construct validity of a writing assessment. In addition, the AAA procedure is considered to be an operable method within a large-scale assessment, while the results for the anchor items as stand-alone units are promising as well.

In conclusion, the results presented in this study show that a rating scale with anchor essays is a useful addition to an analytical rating procedure. Suggestions for future research would be to further explore the sources of differences in inter-rater agreement and generalisability between tasks and between aspects of writing. Moreover, it would be interesting to examine the extent to which a rating scale with anchor essays can be used as a reliable and valid stand-alone evaluation tool for different assessment purposes.

## Appendix A – Example of an average essay with assessment criteria

**(translated from Dutch)**

**Aspect: Content**

**Task: C – Yummy**

3 April 2009

Dear, Smikkel Company.
Sorry, for having enclosed two bars.
You see that is because: It was not 30 April yet.
You see, there were no coupons on the baars any more.
I hope that you, accept it all the same.
My name is Irene van der wagt
I live at numer eight jacobadamsingel
Postcode: 53 md Zutphen

Thanks for reading this,

Best wishes irene

**+/- The student gives all information necessary to receive the telephone**
The following elements are present:
- \- the student mentions that 8 point and 2 wraps are enclosed in the letter
- \+ the student mentions that wraps with points are no more available in shops
- \+ the student mentions that the promotional campaign is still on (it is not yet April 30l)
- \- the student requests Yummy to send the telephone
- \+ the student mentions his address
- +/- the envelope is addressed correctly

**- The student mentions extra information, helping him to receive the telephone.**
Example:
- \- the students mentions extra reasons to receive the telephone (apart from the applicable promotional campaign)

**- It is clear that the letter has been written in order to receive the telephone.**
The letter contains:
- \- convincing arguments to receive the telephone nonetheless
- \+ sufficient information to receive the telephone at home

**+ It is clear that the letter is written to an unknown person at the Yummy company. The student has taken his audience into account**:
- \+ The student addresses a general person.
- \+ The tone of the letter is polite.
- \+ No prior knowledge is assumed (except knowledge about the promotional campaign)

**Summary:**

**A tidy, but incomplete letter: the student does not fully explain the problem and does not explicitly ask for the phone.**

## Appendix B – Selection of anchor essays

| Task A - Contents | | | Task B - Contents | | | Task C - Contents | | |
|---|---|---|---|---|---|---|---|---|
| **mean z-score** | **sd z-score** | **mean scale score** | **mean z-score** | **sd z-score** | **mean scale score** | **mean z-score** | **sd z-score** | **mean scale score** |
| -1.90 | 0.50 | 69 | -2.11 | 0.48 | 64 | -1.46 | 0.17 | 75 |
| -1.85 | 0.70 | 70 | -1.99 | 0.58 | 66 | -1.40 | 0.09 | 76 |
| -1.54 | 0.74 | 75 | -1.76 | 0.27 | 70 | -1.36 | 0.43 | 77 |
| -1.22 | 0.72 | 80 | -1.28 | 0.32 | 78 | -1.25 | 0.38 | 79 |
| -0.94 | 0.29 | 85 | -0.91 | 0.80 | 85 | **-1.19** | **0.21** | **80*** |
| **-0.87** | **0.10** | **86*** | -0.91 | 0.88 | 85 | -0.95 | 0.41 | 84 |
| -0.76 | 0.58 | 88 | **-0.87** | **0.38** | **85*** | -0.93 | 0.26 | 84 |
| -0.72 | 0.33 | 88 | -0.87 | 0.38 | 85 | -0.91 | 0.35 | 85 |
| -0.70 | 0.48 | 89 | -0.59 | 0.62 | 90 | -0.83 | 0.44 | 86 |
| -0.67 | 0.14 | 89 | -0.47 | 0.44 | 92 | -0.69 | 0.53 | 88 |
| -0.65 | 0.17 | 89 | -0.43 | 0.45 | 93 | -0.66 | 0.26 | 89 |
| -0.61 | 0.13 | 90 | -0.40 | 0.35 | 93 | -0.59 | 0.62 | 90 |
| -0.55 | 0.64 | 91 | -0.38 | 0.17 | 93 | -0.54 | 0.53 | 91 |
| -0.43 | 0.48 | 93 | -0.31 | 0.39 | 95 | -0.48 | 0.41 | 92 |
| -0.24 | 0.24 | 96 | -0.28 | 0.31 | 95 | -0.42 | 0.49 | 93 |
| -0.13 | 0.30 | 98 | -0.23 | 0.37 | 96 | -0.35 | 0.66 | 94 |
| -0.12 | 0.23 | 98 | -0.08 | 0.60 | 99 | -0.32 | 0.67 | 95 |
| -0.12 | 0.23 | 98 | -0.02 | 0.72 | 100 | -0.17 | 0.39 | 97 |
| -0.10 | 0.49 | 98 | **0.10** | **0.09** | **102*** | -0.06 | 0.84 | 99 |
| **-0.05** | **0.21** | **99*** | 0.13 | 0.45 | 102 | **-0.05** | **0.19** | **99*** |
| 0.03 | 0.43 | 100 | 0.16 | 0.34 | 103 | 0.06 | 0.64 | 101 |
| 0.07 | 0.33 | 101 | 0.20 | 0.22 | 103 | 0.06 | 0.23 | 101 |
| 0.11 | 0.41 | 102 | 0.21 | 0.45 | 104 | 0.19 | 0.53 | 103 |
| 0.11 | 0.34 | 102 | 0.24 | 0.72 | 104 | 0.20 | 0.39 | 103 |
| 0.16 | 0.23 | 103 | 0.30 | 0.39 | 105 | 0.23 | 0.76 | 104 |
| 0.22 | 0.32 | 103 | 0.42 | 0.33 | 107 | 0.26 | 0.53 | 104 |
| 0.26 | 0.47 | 104 | 0.43 | 0.55 | 107 | 0.39 | 0.64 | 107 |
| 0.26 | 0.25 | 104 | 0.44 | 0.88 | 108 | 0.50 | 0.41 | 108 |
| 0.27 | 0.19 | 104 | 0.47 | 0.65 | 108 | 0.65 | 0.61 | 111 |
| 0.70 | 0.08 | 111 | 0.57 | 0.24 | 110 | 0.68 | 0.76 | 112 |
| 0.77 | 0.42 | 113 | 0.60 | 0.48 | 110 | 0.69 | 0.55 | 112 |
| 0.81 | 0.20 | 113 | 0.69 | 0.31 | 112 | 0.82 | 0.75 | 114 |
| 0.84 | 0.38 | 114 | 0.79 | 0.18 | 113 | 0.91 | 0.42 | 115 |
| 0.88 | 0.27 | 114 | 0.81 | 0.71 | 114 | **0.92** | **0.17** | **116*** |
| **0.91** | **0.15** | **115*** | 0.91 | 0.32 | 115 | 0.96 | 0.17 | 116 |
| 0.96 | 0.31 | 115 | 1.07 | 0.93 | 118 | 1.00 | 0.34 | 117 |
| 1.26 | 0.58 | 120 | **1.23** | **0.15** | **121*** | 1.06 | 0.61 | 118 |
| 1.47 | 0.29 | 124 | 1.27 | 0.35 | 122 | 1.42 | 0.17 | 124 |
| 1.76 | 0.51 | 128 | 1.40 | 0.30 | 124 | 1.80 | 0.75 | 131 |
| 2.34 | 0.55 | 138 | 1.47 | 0.16 | 125 | 1.82 | 0.22 | 131 |

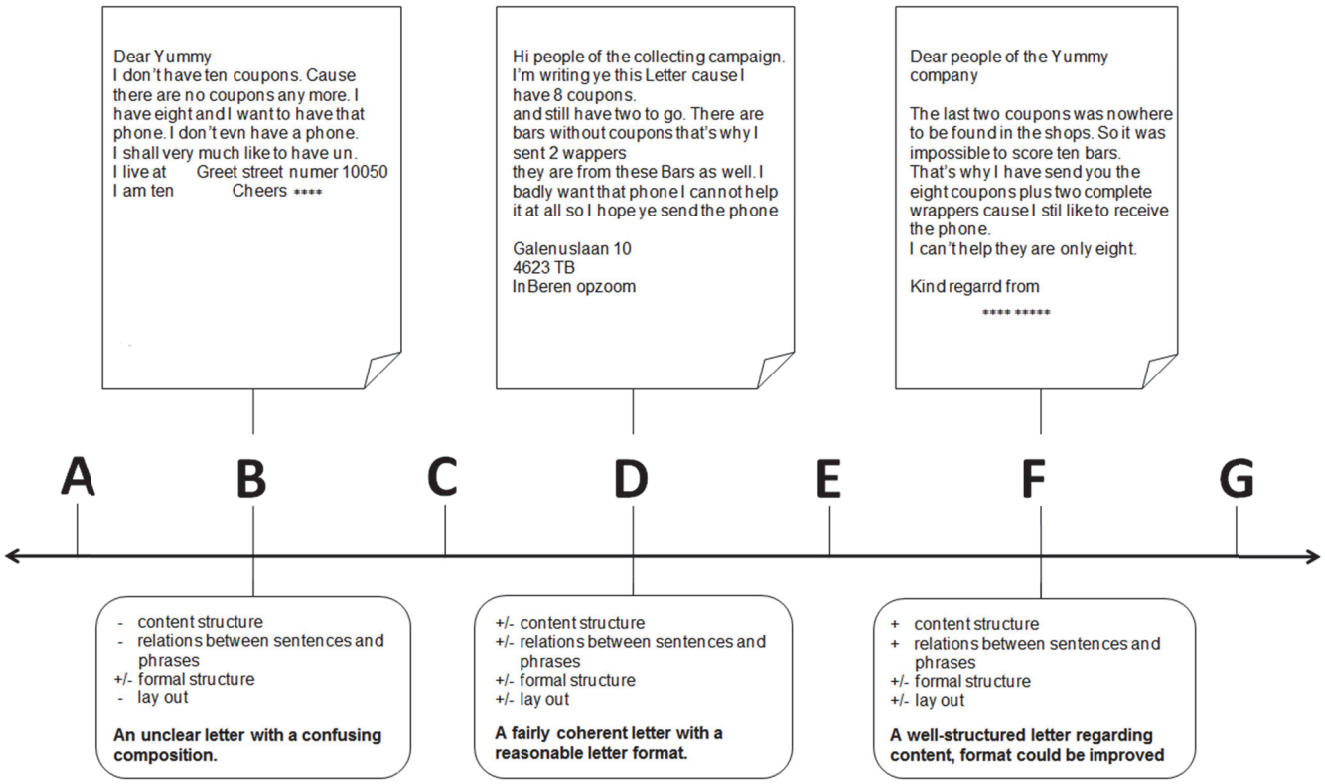
* essay selected as anchor

| Task A – Structure | | | Task B - Structure | | | Task C - Structure | | |
|---|---|---|---|---|---|---|---|---|
| **mean z-score** | **sd z-score** | **mean scale score** | **mean z-score** | **sd z-score** | **mean scale score** | **mean z-score** | **sd z-score** | **mean scale score** |
| -1.65 | 0.32 | 72 | -2.07 | 0.37 | 63 | -1.78 | 0.17 | 70 |
| -1.55 | 0.39 | 74 | -1.90 | 0.40 | 66 | -1.72 | 0.68 | 71 |
| -1.23 | 0.62 | 79 | -1.65 | 0.36 | 71 | -1.38 | 0.32 | 77 |
| -1.09 | 0.21 | 82 | -1.34 | 0.29 | 76 | -1.35 | 0.34 | 77 |
| -0.92 | 0.60 | 84 | -0.92 | 0.60 | 84 | -1.13 | 0.54 | 81 |
| **-0.90** | **0.20** | **85*** | -0.84 | 0.44 | 85 | -1.08 | 0.70 | 82 |
| -0.85 | 0.65 | 86 | -0.73 | 0.47 | 87 | **-0.85** | **0.41** | **86*** |
| -0.80 | 0.26 | 86 | **-0.69** | **0.28** | **88*** | -0.85 | 0.29 | 86 |
| -0.72 | 0.15 | 88 | -0.51 | 0.43 | 91 | -0.70 | 0.67 | 88 |
| -0.67 | 0.32 | 89 | -0.49 | 0.14 | 91 | -0.62 | 0.41 | 89 |
| -0.56 | 0.19 | 91 | -0.31 | 0.28 | 94 | -0.58 | 0.74 | 90 |
| -0.55 | 0.22 | 91 | -0.27 | 0.37 | 95 | -0.43 | 0.36 | 93 |
| -0.50 | 0.38 | 92 | -0.25 | 0.81 | 96 | -0.44 | 0.42 | 93 |
| -0.43 | 0.25 | 93 | -0.15 | 1.33 | 97 | -0.37 | 0.50 | 94 |
| -0.40 | 0.46 | 93 | -0.17 | 0.59 | 97 | -0.33 | 0.38 | 94 |
| -0.40 | 0.46 | 93 | -0.18 | 0.62 | 97 | -0.35 | 0.33 | 94 |
| -0.36 | 0.17 | 94 | -0.15 | 0.54 | 97 | -0.29 | 0.52 | 95 |
| -0.32 | 0.27 | 95 | -0.12 | 0.84 | 98 | -0.09 | 0.56 | 98 |
| -0.32 | 0.71 | 95 | **-0.03** | **0.27** | **99*** | -0.06 | 0.45 | 99 |
| -0.25 | 0.58 | 96 | -0.06 | 0.82 | 99 | **-0.01** | **0.45** | **100*** |
| -0.20 | 0.43 | 97 | 0.03 | 0.51 | 101 | 0.08 | 0.67 | 101 |
| -0.11 | 0.38 | 98 | 0.06 | 0.67 | 101 | 0.07 | 0.47 | 101 |
| 0.00 | 0.75 | 100 | 0.24 | 0.71 | 104 | 0.08 | 0.58 | 101 |
| **0.13** | **0.37** | **102*** | 0.20 | 0.67 | 104 | 0.19 | 0.53 | 103 |
| 0.20 | 0.46 | 103 | 0.31 | 0.34 | 105 | 0.38 | 0.52 | 107 |
| 0.34 | 0.36 | 106 | 0.29 | 0.46 | 105 | 0.44 | 0.31 | 107 |
| 0.49 | 0.60 | 108 | 0.36 | 0.67 | 106 | 0.49 | 0.68 | 108 |
| 0.45 | 0.48 | 108 | 0.33 | 0.39 | 106 | 0.44 | 0.62 | 108 |
| 0.58 | 0.64 | 110 | 0.49 | 0.49 | 109 | 0.58 | 0.52 | 110 |
| 0.67 | 0.94 | 111 | 0.50 | 0.58 | 109 | 0.68 | 0.43 | 111 |
| 0.65 | 0.48 | 111 | 0.69 | 0.36 | 112 | 0.75 | 0.39 | 113 |
| **0.76** | **0.37** | **113*** | 0.75 | 0.63 | 113 | 0.84 | 0.35 | 114 |
| 0.79 | 0.49 | 113 | 0.81 | 0.92 | 114 | 0.81 | 0.40 | 114 |
| 1.06 | 0.44 | 118 | **0.82** | **0.20** | **115*** | **0.82** | **0.42** | **114*** |
| 1.08 | 0.48 | 118 | 0.89 | 0.51 | 116 | 0.86 | 0.18 | 115 |
| 1.26 | 0.46 | 121 | 0.94 | 0.12 | 117 | 0.90 | 0.46 | 115 |
| 1.29 | 0.11 | 122 | 0.94 | 0.32 | 117 | 1.19 | 0.55 | 120 |
| 1.61 | 0.27 | 127 | 1.25 | 0.48 | 122 | 1.54 | 0.35 | 126 |
| 1.65 | 1.09 | 128 | 1.25 | 0.35 | 122 | 1.60 | 0.28 | 127 |
| 1.75 | 0.48 | 130 | 1.68 | 0.58 | 130 | 1.64 | 0.64 | 128 |

| Task A - Correctness | | | Task B - Correctness | | | Task C - Correctness | | |
|---|---|---|---|---|---|---|---|---|
| **mean z-score** | **sd z-score** | **mean scale score** | **mean z-score** | **sd z-score** | **mean scale score** | **mean z-score** | **sd z-score** | **mean scale score** |
| -2.26 | 0.28 | 62 | -1.87 | 0.21 | 66 | -1.64 | 0.25 | 70 |
| -2.01 | 0.43 | 66 | -1.53 | 0.45 | 72 | -1.15 | 0.39 | 79 |
| -1.47 | 0.51 | 75 | -1.50 | 0.49 | 73 | -1.13 | 0.29 | 80 |
| -1.00 | 0.85 | 83 | -1.37 | 0.50 | 75 | **-1.03** | **0.10** | **81*** |
| -0.98 | 0.44 | 83 | -1.26 | 0.48 | 77 | -0.99 | 0.65 | 82 |
| -0.96 | 0.75 | 84 | -0.98 | 0.82 | 82 | -0.93 | 0.78 | 83 |
| -0.96 | 0.54 | 84 | -0.90 | 0.62 | 84 | -0.91 | 0.58 | 84 |
| **-0.85** | **0.35** | **86*** | -0.80 | 0.64 | 85 | -0.87 | 0.62 | 84 |
| -0.81 | 0.55 | 86 | **-0.52** | **0.29** | **90*** | -0.81 | 0.75 | 85 |
| -0.67 | 0.79 | 89 | -0.52 | 0.29 | 90 | -0.82 | 0.60 | 85 |
| -0.48 | 0.33 | 92 | -0.42 | 0.50 | 92 | -0.67 | 0.63 | 88 |
| -0.30 | 0.43 | 95 | -0.34 | 1.00 | 94 | -0.53 | 0.39 | 90 |
| -0.28 | 0.31 | 95 | -0.20 | 0.36 | 96 | -0.38 | 0.64 | 93 |
| -0.20 | 0.31 | 97 | -0.21 | 0.37 | 96 | -0.33 | 0.61 | 94 |
| -0.11 | 0.51 | 98 | -0.13 | 0.48 | 98 | -0.21 | 0.73 | 96 |
| **-0.08** | **0.37** | **99*** | -0.08 | 0.94 | 99 | -0.21 | 0.83 | 96 |
| -0.04 | 0.58 | 99 | **-0.08** | **0.37** | **99*** | -0.19 | 0.42 | 97 |
| 0.01 | 0.52 | 100 | -0.03 | 0.47 | 99 | -0.03 | 0.65 | 99 |
| 0.23 | 0.57 | 104 | 0.00 | 0.55 | 100 | -0.08 | 0.57 | 99 |
| 0.22 | 0.38 | 104 | -0.02 | 0.81 | 100 | -0.06 | 0.66 | 99 |
| 0.30 | 0.59 | 105 | 0.03 | 0.41 | 101 | -0.02 | 0.63 | 100 |
| 0.31 | 0.55 | 105 | 0.17 | 0.57 | 103 | 0.14 | 0.33 | 102 |
| 0.38 | 0.21 | 106 | 0.25 | 0.44 | 105 | 0.09 | 0.74 | 102 |
| 0.34 | 0.17 | 106 | 0.26 | 0.21 | 105 | 0.11 | 0.46 | 102 |
| 0.41 | 0.39 | 107 | 0.32 | 0.79 | 106 | 0.10 | 0.83 | 102 |
| 0.47 | 0.43 | 108 | 0.35 | 1.03 | 106 | **0.19** | **0.27** | **103*** |
| 0.44 | 0.38 | 108 | 0.39 | 0.31 | 107 | 0.24 | 0.48 | 104 |
| 0.54 | 0.45 | 109 | 0.43 | 0.70 | 108 | 0.52 | 0.61 | 109 |
| 0.67 | 0.41 | 111 | 0.45 | 0.33 | 108 | 0.59 | 0.67 | 111 |
| 0.71 | 0.66 | 112 | 0.50 | 0.56 | 109 | 0.64 | 0.42 | 112 |
| 0.78 | 0.47 | 113 | 0.55 | 0.79 | 110 | 0.69 | 0.70 | 113 |
| **0.80** | **0.22** | **114*** | **0.63** | **0.47** | **112*** | 0.78 | 0.51 | 114 |
| 0.88 | 0.31 | 115 | 0.73 | 0.57 | 113 | 0.82 | 0.80 | 115 |
| 0.89 | 0.55 | 115 | 0.74 | 0.69 | 114 | **0.94** | **0.36** | **117*** |
| 0.90 | 0.34 | 115 | 0.94 | 0.70 | 117 | 1.18 | 0.43 | 121 |
| 0.95 | 0.19 | 116 | 0.99 | 0.67 | 118 | 1.19 | 0.56 | 122 |
| 0.96 | 0.35 | 116 | 1.17 | 0.89 | 121 | 1.30 | 0.75 | 124 |
| 1.28 | 0.51 | 122 | 1.15 | 0.27 | 121 | 1.60 | 0.49 | 129 |
| 1.75 | 0.57 | 130 | 1.19 | 0.25 | 122 | 1.67 | 0.34 | 130 |
| - | - | - | 1.51 | 0.63 | 128 | - | - | - |

## Appendix C – Rating scale with anchor essays

**C – Structure**



# Appendix D – Rating design

| | A-T&G | | | | | | B-Pookie | | | | | | C-Yummy | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A1** | **A2** | **A3** | **A4** | **A5** | **A6** | **B1** | **B2** | **B3** | **B4** | **B5** | **B6** | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** | **batch** |
| **rater** | | | | | | | | | | | | | | | | | | | essays |
| 1 | 50 | | | | | | | | 50 | | | | | | 50 | | | | 150 |
| 2 | | 50 | | | | | | 50 | | | | | | | | 50 | | | 150 |
| 3 | | | 50 | | | | 50 | | | | | | | | | | 50 | | 150 |
| 4 | | | | 50 | | | | | | 50 | | | | | | | | 50 | 150 |
| 5 | | | | | 50 | | | | | | 50 | | 50 | | | | | | 150 |
| 6 | | | | | | 50 | | | | | | 50 | 50 | | | | | | 150 |
| 7 | | | | | | 50 | 50 | | | | | | | 50 | | | | | 150 |
| 8 | | | | | 50 | | | 50 | | | | | | | | 50 | | | 150 |
| 9 | | | | 50 | | | 50 | | | | | | | | | | 50 | | 150 |
| 10 | | | 50 | | | | | | | 50 | | | 50 | | | | | | 150 |
| 11 | | 50 | | | | | | | | | 50 | | | | | | | 50 | 150 |
| 12 | 50 | | | | | | | | | | | 50 | | 50 | | | | | 150 |
| 13 | 50 | | | | | | | | 50 | | | | | | 50 | | | | 150 |
| **ratings** | **3** | **2** | **2** | **2** | **2** | **2** | **3** | **2** | **2** | **2** | **2** | **2** | **3** | **2** | **2** | **2** | **2** | **2** | |

| | A-T&G | | | | | | B-Pookie | | | | | | C-Yummy | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **A1** | **A2** | **A3** | **A4** | **A5** | **A6** | **B1** | **B2** | **B3** | **B4** | **B5** | **B6** | **C1** | **C2** | **C3** | **C4** | **C5** | **C6** | **batch** |
| **rater** | | | | | | | | | | | | | | | | | | | essays |
| 14 | 50 | | | | | | | | 50 | | | | | | 50 | | | | 150 |
| 15 | | 50 | | | | | | 50 | | | | | | | | 50 | | | 150 |
| 16 | | | 50 | | | | 50 | | | | | | | | | | 50 | | 150 |
| 17 | | | | 50 | | | | | | 50 | | | | | | | | 50 | 150 |
| 18 | | | | | 50 | | | | | | 50 | | 50 | | | | | | 150 |
| 19 | | | | | | 50 | | | | | | 50 | 50 | | | | | | 150 |
| 20 | | | | | | 50 | 50 | | | | | | | 50 | | | | | 150 |
| 21 | | | | | 50 | | | 50 | | | | | | | | 50 | | | 150 |
| 22 | | | | 50 | | | 50 | | | | | | | | | | 50 | | 150 |
| 23 | | | 50 | | | | | | | 50 | | | 50 | | | | | | 150 |
| 24 | | 50 | | | | | | | | | 50 | | | | | | | 50 | 150 |
| 25 | 50 | | | | | | | | | | | 50 | | 50 | | | | | 150 |
| 26 | 50 | | | | | | | | 50 | | | | | | 50 | | | | 150 |
| **ratings** | **3** | **2** | **2** | **2** | **2** | **2** | **3** | **2** | **2** | **2** | **2** | **2** | **3** | **2** | **2** | **2** | **2** | **2** | |

## Appendix E – Variance components Condition 2

Variance components of observed score for Condition 2 (new, analytical questions only)

| Aspect | Content | Structure | Correctness |
|---|---|---|---|
| **Variance component** | **%** | **%** | **%** |
| Person | 26.80 | 37.00 | 22.60 |
| Task | 3.00 | 5.90 | 0.00 |
| Rater | 0.00 | 0.00 | 0.70 |
| Person*Task | 36.90 | 21.80 | 4.00 |
| Person*Rater | 0.00 | 0.00 | 0.00 |
| Task*Rater | 0.60 | 0.60 | 0.10 |
| Random error | 32.60 | 34.70 | 72.70 |

Variance components of observed score for Condition 2 (new, anchor items only)

| Aspect | Content | Structure | Correctness |
|---|---|---|---|
| **Variance component** | **%** | **%** | **%** |
| Person | 29.20 | 38.90 | 48.00 |
| Task | 2.00 | 0.00 | 0.00 |
| Rater | 1.90 | 0.00 | 0.80 |
| Person*Task | 25.10 | 13.60 | 10.30 |
| Person*Rater | 2.10 | 7.70 | 1.00 |
| Task*Rater | 1.30 | 0.90 | 0.30 |
| Random error | 38.50 | 39.00 | 39.40 |

## Appendix F –Test statistics for all Grades (3 to 6)

Test statistics per Task, per Condition (averages over candidates and raters)

| Task A | Condition 1 (old) | p-value[1] | score[2] | sd[3] | Rit[4] | Alpha[5] (*) | Condition 2 (new) | p-value | score | sd | Rit | Alpha (*) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Total test** (22 items) | **36.6** | **8.79** | **4.02** | **0.44** | **0.79 (0.87)** | **Total test** (31 items) | **47.4** | **21.8** | **10.55** | **0.57** | **0.91 (0.93)** |
| | **Content** (17 items) | 31.4 | 5.34 | 2.63 | 0.43 | 0.73 (0.86) | **Content** (17 items) | 45.8 | 10.1 | 5.23 | 0.58 | 0.85 (0.93) |
| | **Structure** (3 items) | 57.5 | 1.72 | 0.90 | 0.68 | 0.39 (0.90) | **Structure** (10 items) | 49.1 | 7.4 | 3.84 | 0.66 | 0.79 (0.94) |
| | **Correctness** (2 items) | 43.0 | 1.72 | 1.31 | 0.83 | 0.54 (0.96) | **Correctness** (4 items) | 48.4 | 4.4 | 2.73 | 0.86 | 0.72 (0.96) |
| **Task B** | **Condition 1 (old)** | **p-value** | **score** | **sd** | **Rit** | **Alpha (*)** | **Condition 2 (new)** | **p-value** | **score** | **sd** | **Rit** | **Alpha (*)** |
| | **Total test** (22 items) | **51.7** | **13.45** | **4.62** | **0.47** | **0.82 (0.90)** | **Total test** (28 items) | **57.1** | **24.5** | **8.37** | **0.52** | **0.88 (0.91)** |
| | **Content** (16 items) | 54.7 | 8.76 | 3.15 | 0.49 | 0.79 (0.90) | **Content** (13 items) | 61.8 | 11.12 | 3.26 | 0.54 | 0.73 (0.89) |
| | **Structure** (4 items) | 37.5 | 2.25 | 1.18 | 0.63 | 0.48 (0.90) | **Structure** (11 items) | 56.0 | 8.96 | 3.83 | 0.60 | 0.77 (0.93) |
| | **Correctness** (2 items) | 61.1 | 2.44 | 1.29 | 0.90 | 0.77 (0.99) | **Correctness** (4 items) | 49.6 | 4.47 | 2.55 | 0.84 | 0.71 (0.96) |
| **Task C** | **Condition 1 (old)** | **p-value** | **score** | **sd** | **Rit** | **Alpha (*)** | **Condition 2 (new)** | **p-value** | **score** | **sd** | **Rit** | **Alpha (*)** |
| | **Total test** (21 items) | **47.9** | **11.02** | **5.72** | **0.56** | **0.88 (0.94)** | **Total test** (30 items) | **50.2** | **22.6** | **9.10** | **0.49** | **0.88 (0.91)** |
| | **Content** (13 items) | 48.6 | 6.32 | 3.3 | 0.57 | 0.82 (0.93) | **Content** (14 items) | 51.1 | 9.71 | 3.80 | 0.49 | 0.73 (0.88) |
| | **Structure** (6 items) | 44.0 | 2.64 | 1.69 | 0.63 | 0.70 (0.94) | **Structure** (12 items) | 51.8 | 8.80 | 4.17 | 0.61 | 0.81 (0.93) |
| | **Correctness** (2 items) | 51.5 | 2.06 | 1.41 | 0.91 | 0.79 (0.99) | **Correctness** (4 items) | 45.1 | 4.06 | 2.63 | 0.83 | 0.69 (0.96) |

* Coefficient Alpha for test length of 40 items (using the Spearman-Brown prophecy formula)

[1] p-value: score on a percentage scale
[2] score: number of score points
[3] sd: standard deviation of score
[4] Rit: average discrimination index (correlation between items and complete test)
[5] coefficient Alpha: measure for test reliability

## Appendix G – Test statistics per Grade

Test statistics for all Tasks (A-B-C) (averages over candidates and raters)

| **Condition 1 (old)** | **p-value** | **score** | **sd** | **Rit** | **Alpha** | **Condition 2 (new)** | **p-value** | **score** | **sd** | **Rit** | **Alpha** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade 3 | 35.02 | 7.39 | 5.04 | 0.52 | 0.86 | Grade 3 | 40.47 | 18.01 | 9.08 | 0.53 | 0.89 |
| Grade 4 | 45.08 | 11.00 | 4.37 | 0.45 | 0.80 | Grade 4 | 48.81 | 21.77 | 9.12 | 0.52 | 0.88 |
| Grade 5 | 48.62 | 11.86 | 4.22 | 0.44 | 0.79 | Grade 5 | 55.82 | 24.88 | 8.28 | 0.48 | 0.86 |
| Grade 6 | 53.23 | 12.98 | 4.33 | 0.46 | 0.79 | Grade 6 | 61.63 | 27.49 | 8.05 | 0.49 | 0.86 |

Test statistics for Task A (Narrative) (averages over candidates and raters)

| **Condition 1 (old)** | **p-value** | **score** | **sd** | **Rit** | **Alpha** | **Condition 2 (new)** | **p-value** | **score** | **sd** | **Rit** | **Alpha** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade 3 | 29.73 | 7.13 | 3.92 | 0.45 | 0.81 | Grade 3 | 35.3 | 16.2 | 9.85 | 0.55 | 0.91 |
| Grade 4 | 37.73 | 9.06 | 3.74 | 0.41 | 0.76 | Grade 4 | 45.6 | 21.0 | 9.96 | 0.54 | 0.90 |
| Grade 5 | 37.43 | 8.98 | 4.05 | 0.43 | 0.79 | Grade 5 | 50.7 | 23.3 | 10.16 | 0.56 | 0.91 |
| Grade 6 | 41.88 | 10.05 | 3.88 | 0.43 | 0.75 | Grade 6 | 59.0 | 27.2 | 9.27 | 0.54 | 0.89 |

Test statistics for Task B (Directive) (averages over candidates and raters)

| **Condition 1 (old)** | **p-value** | **score** | **sd** | **Rit** | **Alpha** | **Condition 2 (new)** | **p-value** | **score** | **sd** | **Rit** | **Alpha** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade 3 | 42.11 | 10.95 | 5.65 | 0.54 | 0.88 | Grade 3 | 48.5 | 20.9 | 8.77 | 0.54 | 0.89 |
| Grade 4 | 50.02 | 13.01 | 4.04 | 0.41 | 0.77 | Grade 4 | 52.4 | 22.6 | 8.29 | 0.51 | 0.87 |
| Grade 5 | 55.50 | 14.43 | 3.58 | 0.38 | 0.72 | Grade 5 | 61.4 | 26.4 | 7.10 | 0.45 | 0.84 |
| Grade 6 | 59.68 | 15.52 | 3.72 | 0.41 | 0.76 | Grade 6 | 66.7 | 28.7 | 6.85 | 0.47 | 0.84 |

Test statistics for Task C (Argumentative) (averages over candidates and raters)

| Condition 1 (old) | p-value | score | sd | Rit | Alpha | Condition 2 (new) | p-value | score | sd | Rit | Alpha |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Grade 3 | 33.22 | 7.64 | 5.56 | 0.57 | 0.89 | Grade 3 | 37.6 | 16.9 | 8.63 | 0.5 | 0.88 |
| Grade 4 | 47.49 | 10.92 | 5.33 | 0.53 | 0.86 | Grade 4 | 48.4 | 21.8 | 9.11 | 0.50 | 0.88 |
| Grade 5 | 52.93 | 12.17 | 5.03 | 0.51 | 0.85 | Grade 5 | 55.4 | 24.9 | 7.59 | 0.42 | 0.83 |
| Grade 6 | 58.13 | 13.37 | 5.38 | 0.54 | 0.87 | Grade 6 | 59.2 | 26.6 | 8.04 | 0.46 | 0.85 |

# Appendix H – Correlations between scores per Aspect and Task

Average correlationsa across all grades, Condition 1 (old)

| | | Task A | | | Task B | | | Task C | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Aspect** | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **2** | **3** | **Reliability[b]** |
| A | 1 – Content | 1.00 | | | | | | | | | 0.54 |
| | 2 - Structure | 1.00 | 1.00 | | | | | | | | 0.55 |
| | 3 - Correctness | 1.00 | 1.00 | 1.00 | | | | | | | 0.16 |
| B | 1 – Content | **0.53** | 0.51 | 0.84 | 1.00 | | | | | | 0.89 |
| | 2 – Structure | 0.63 | **0.76** | 1.00 | 0.76 | 1.00 | | | | | 0.72 |
| | 3 - Correctness | 0.66 | 1.22 | **0.09** | 0.55 | 1.00 | 1.00 | | | | 0.25 |
| C | 1 – Content | **0.56** | 0.59 | 0.59 | **0.42** | 0.31 | 0.24 | 1.00 | 0.77 | 0.15 | 0.77 |
| | 2 – Structure | 1.00 | **0.87** | 1.37 | 0.47 | **0.60** | 0.56 | 0.84 | 1.00 | 0.76 | 0.76 |
| | 3 - Correctness | 1.00 | 1.00 | **1.00** | 0.87 | 1.00 | **1.00** | 1.00 | 1.00 | 1.00 | 0.15 |

[a] Corrected for attenuation
[b] Spearman Brown

Average correlationsa across all grades, Condition 2 (new) (analytical questions only)

| | | Task A | | | Task B | | | Task C | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Aspect** | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **2** | **3** | **Reliability[b]** |
| A | 1 - Content | 1.00 | | | | | | | | | 0.782 |
| | 2 - Structure | 1.00 | 1.00 | | | | | | | | 0.795 |
| | 3 - Correctness | 1.00 | 1.00 | 1.00 | | | | | | | 0.159 |
| B | 1 - Content | **0.48** | 0.46 | 0.91 | 1.00 | | | | | | 0.756 |
| | 2 - Structure | 0.51 | **0.61** | 1.00 | 0.80 | 1.00 | | | | | 0.765 |
| | 3 - Correctness | 0.49 | 0.92 | **0.08** | 0.54 | 1.00 | 1.00 | | | | 0.307 |
| C | 1 - Content | **0.44** | 0.47 | 0.56 | **0.44** | 0.28 | 0.21 | 1.00 | | | 0.857 |
| | 2 - Structure | 0.71 | **0.73** | 1.37 | 0.51 | **0.59** | 0.51 | 0.91 | 1.00 | | 0.761 |
| | 3 - Correctness | 0.42 | 0.59 | **0.91** | 0.47 | 0.49 | **0.51** | 0.56 | 0.87 | 1.00 | 0.584 |

[a] Corrected for attenuation
[b] Spearman Brown

Average correlationsa across all grades, Condition 2 (new) (anchor questions only)

| | | Task A | | | Task B | | | Task C | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Aspect** | **1** | **2** | **3** | **1** | **2** | **3** | **1** | **2** | **3** | **Reliability[b]** |
| A | 1 - Content | 1.00 | | | | | | | | | 0.734 |
| | 2 - Structure | 1.00 | 1.00 | | | | | | | | 0.776 |
| | 3 - Correctness | 0.97 | 1.00 | 1.00 | | | | | | | 0.759 |
| B | 1 - Content | **0.68** | 0.62 | 0.75 | 1.00 | | | | | | 0.574 |
| | 2 - Structure | 0.87 | **0.97** | 1.00 | 1.00 | 1.00 | | | | | 0.569 |
| | 3 - Correctness | 0.77 | 0.86 | **0.83** | 1.00 | 1.17 | 1.00 | | | | 0.722 |
| C | 1 - Content | **0.59** | 0.58 | 0.52 | **0.47** | 0.65 | 0.59 | 1.00 | | | 0.840 |
| | 2 - Structure | 0.65 | **0.71** | 0.65 | 0.60 | **0.86** | 0.78 | 1.24 | 1.00 | | 0.737 |
| | 3 - Correctness | 0.65 | 0.77 | **0.81** | 0.64 | 0.80 | **0.87** | 0.88 | 1.00 | 1.00 | 0.764 |

[a] Corrected for attenuation
[b] Spearman Brown

# 3 Evaluating text quality: The validity of a revision test for novice writers



## 3.1 Introduction

The complex cognitive task of producing a text involves a range of activities that are generally grouped into three components: *planning* what to write, *translating* these thoughts into words and *reviewing* the text hitherto produced (Flower & Hayes, 1981; Kellogg, 1996; Chenoweth & Hayes, 2003). In a recent model of the writing process, Hayes (2012) added the component of transcribing words, which follows the translation of thoughts. Figure 1 is a schematic representation of the four components of the writing process, based on Hayes' model. Instead of referring to parts of the writing process that are carried out in linear order, these components represent the recursive cognitive processes that comprise writing, all of which can occur at any moment during this activity (Galbraith, 2009).

Bereiter and Scardamalia (1987) described the development of these cognitive processes in three stages of increasing sophistication: from knowledge *telling* via knowledge *transforming* to knowledge *crafting*. In the first stage, (novice) writers are mainly concerned with restating their thoughts by generating textual output, and reviewing is limited to minimal lexical changes. In the transitional stage, the interaction between the writer's ideas and the generated text is established, and reviewing can lead to additional planning or text production. In the final stage, (skilled) writers shape their text with the potential reader in mind and continue to re-read, evaluate and improve their texts (Bereiter & Scaradamalia, 1987; Kellogg, 2008).

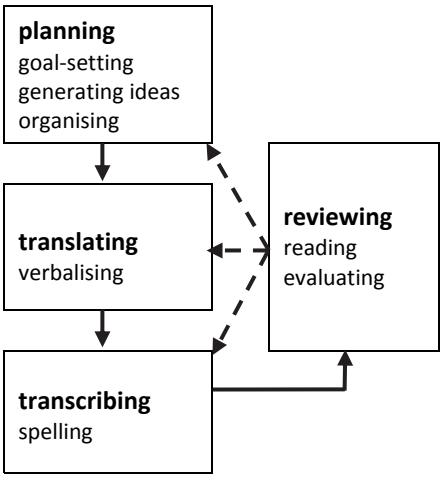



*Figure 1*. A schematic representation of the writing process.
(based on Flower & Hayes, 1981; Hayes, 2012)

Writing, in other words, involves reviewing and rewriting ones text, with the targeted reader in mind. This characteristic distinguishes the production of written text from the production of spoken text (Sommers, 1980). Whereas speaking is typically an interactive process in which the sender of the message can confirm understanding, and the recipient can ask for clarification, the process of writing lacks this instantaneous feedback. Therefore, writers have to anticipate the reader's reception of the language they produce in order to communicate effectively (Van der Pool, 1995). Taking the above into account, reviewing in combination with rewriting—that is, *revising*—is a substantial part of the cognitive writing process, and the ability to revise one's text is one of the characteristics that define a skilled writer.

Supporting this notion, several studies found a positive relation between the occurrence of text revision and text quality. The results of the US National Assessment for Educational Progress (NAEP) showed that eleventh graders who reported that they revised their texts frequently outperformed students who reported that they never or hardly ever revised (Mullis et al., 1994). A positive relation between the occurrence of text revision and text quality was also found in experimental studies by Stevenson, Schoonen, and De Glopper (2006) and Groenendijk, Janssen, Rijlaarsdam, and Van den Bergh (2008). Furthermore, Barkaoui (2007) summarized findings from research on revision skills, showing that skilled writers differ from less skilled writers in that the former are concerned with their audience during revision. They also revise more frequently, in more stages in the writing process, at multiple text levels, and for multiple purposes. Finally, they actually improve their texts by revising them.

In light of previous research, the ability to revise is an important skill both to acquire and to monitor (Breland, 1999; Rijlaarsdam et al., 2012). In practice, however, textual revision receives little attention in language classes, and students scarcely revise their texts

during writing lessons (Van Gelderen, 1997). However, when they are explicitly asked to detect and solve problems in their texts, students prove to be capable of successful revision (Bereiter & Scardamalia, 1987), which indicates that students do possess the skills that are needed to revise a text.

Insight into the ability of students to revise their texts on the one hand and ways that test makers could improve the content validity of writing assessments on the other hand would be a valuable addition to the evaluation of writing ability. Indeed, the latter could incorporate the assessment of revision skills. The aim of the present study is to investigate how a revision test can be validly and reliably incorporated into a large-scale assessment of writing ability in primary education. In Section 3.2, the construct of text revision is investigated, and different formats for revision tests are discussed. The validity of a multiple-choice test of revision ability is evaluated in Section 3.3, followed by the evaluation of a constructed-response revision test in Section 3.4. Section 3.5 discusses the results of both test formats, which leads to a concluding discussion of the validity of revision tests for novice writers.

## 3.2 Assessing text revision

### 3.2.1 Introduction

A revision test typically consists of a set of multiple-choice questions concerning a text that contains flaws. The test taker corrects these flaws by choosing the right alternatives. In this form, revision tests have a long tradition of serving as a so-called *indirect* measurement of writing, that is, they are aimed to evaluate writing ability indirectly by assessing a related skill. *Direct* writing assessments, on the other hand, usually consist of a writing task designed to elicit a written product that is rated using a scoring instruction. However, productive language skills are notoriously difficult to assess reliably and validly (cf. Godshalk, Swineford, & Coffman, 1966; Breland, Camp, Jones, Morris & Rock, 1987; Cushing Weigle, 2002; Knoch, 2011). Apart from task effects, the scores given by different raters usually diverge, as do the scores given by the same rater when asked to re-evaluate the essay (Van den Bergh & Eiting, 1989; Schoonen, 2005). Hence, to eliminate the effects of task and rater, multiple tasks and multiple raters are needed (Meuffels, 1994; Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012), which would require increased financial and human resources. Thus, indirect assessments of writing have been developed to overcome these practical issues.

The validity of indirect writing assessments is supported by the fact that text revision includes editing on all levels of text (Flower & Hayes, 1981; Hayes, Flower, Schriver, Stratman, & Carey, 1987). Hence, all components of writing are covered by the activity of revising. In order to validate the use of revision tests as an indirect measure of writing ability, various studies have been conducted on the predictive power of indirect writing tests. In past decades, this was usually done by determining the correlation between the results of direct and indirect assessments of writing ability. Godshalk, Swineford and Coffman (1966) investigated the validity of different approaches to the measurement of writing skills, developed at Educational Testing Service and used in secondary schools. Wesdorp (1974) studied the indirect assessment of the writing of fifth-grade pupils in Dutch primary schools. Wesdorp compared the scores on objective tests to the scores on five different essays and found fairly high correlations (± .70). Breland and Gaynor (1979) compared the scores on three different essays with results of the Test of Standard Written English and found correlations varying from .56 to .74. Breland, Camp, Jones, Morris, and Rock (1987) conducted an investigation similar to Godshalk et al. (1966), using data collected from post-secondary pupils. A correlation from .56 to .66 was reported in this study. Table 1 provides an overview of the studies discussed above.

Table 1. *Studies on the Validity of Indirect Measurements of Writing Ability*

| Study (year) | Age pupils | Number of pupils | Correlation indirect and direct measure |
|---|---|---|---|
| Godshalk et al. (1966) | 16-17 | 646 | .71 -.77 |
| Wesdorp (1974) | 12 | 213 | .67 -.68 |
| Breland & Gaynor (1979) | 18 | 234 - 926 | .56 -.74 |
| Breland et al. (1987) | > 18 | 267 | .56 -.66 |

Although the studies summarized in Table 1 show that revision tests can predict actual writing ability fairly well, the poor face validity of this method has led to its diminishing use as a substitute for direct writing assessment. Because writing is a productive language ability, it is often considered invalid to use a multiple-choice test to assess this ability indirectly (cf. Pullens, Den Ouden, Herrlitz, & Van den Bergh, 2013). Consequently, revision tests nowadays are seldom deployed as a predictor of writing ability and a substitute for direct writing assessment. Instead, they are administered as part of a writing assessment and are specifically aimed at the evaluation of the ability to revise a text.

### 3.2.2 The construct of text revision

Text production requires the writer to make a multitude of choices. These decisions can be classified into decisions on *what* to express and decisions on *how* to express it (Inui, Tokunaga, & Tanaka, 1992). The first class comprises the selection and organization of content elements, and the second class covers grammatical and lexical choices. Levelt's model of spoken language production shows a similar classification of activities, referred to as the "conceptualizer" and "formulator" components (Levelt, 1989). Text revision is generally referred to as any instant during the production of a written text in which the writer rereads and—if necessary—edits the text hitherto produced (Flower & Hayes, 1981; Van der Pool, 1995). Revisions can occur at all times during the writing process and can be executed on all text levels, ranging from a change in writing goal to a change in wording or spelling. When revising, a writer ideally reconsiders all the decisions made while writing, that is, those concerning the text as a whole (e.g., writing goal, macro level) and those concerning local features (e.g., wording, micro level). Hence, two types of revision activities can be distinguished: *conceptual* activities on one hand, and *textual* activities on the other hand.

Hayes et al. (1987) listed a variety of conceptual and textual problems that can be detected and corrected during the revision of a text. Figure 2 is a schematic representation of the writing process in which these elements of revision, grouped into corrections at both conceptual and textual levels, are listed. In mapping the revision elements to the components of the writing process, Figure 2 shows that reviewing at the conceptual level can be explained as a reconsideration of the planning activities. Revisions at the conceptual level will therefore concern setting a goal for the text and organizing it, hence resulting in changes in overall structure and content (i.e., changes at the macro/meso level). Similarly, revision at

the textual level concerns the translation of thought into text and is reflected in changes to grammar, wording and spelling (i.e., changes at the micro level).

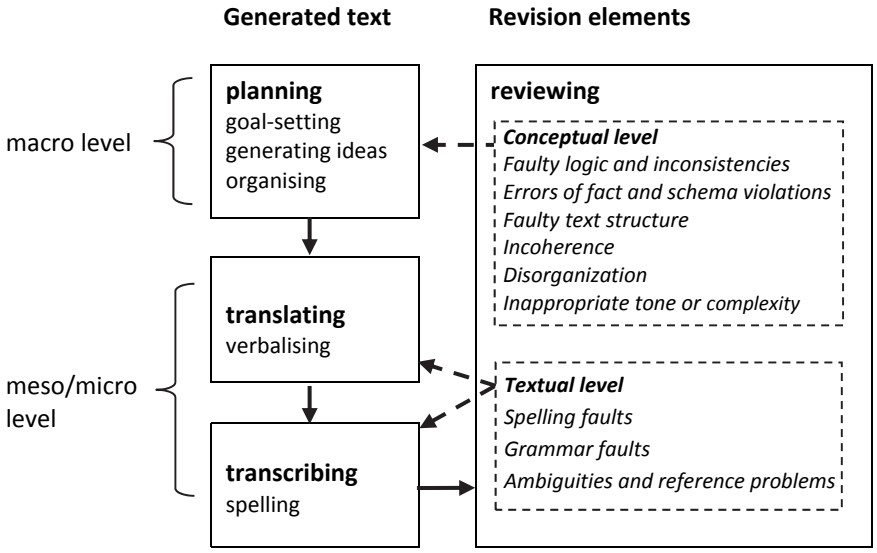



*Figure2*. Elements of the revision process mapped to components of the writing process. (based on Flower & Hayes, 1981; Hayes et al., 1987; Hayes, 2012)

### 3.2.3 Text formats

The construction of a test on revision ability requires several decisions regarding the test format, which should take into account both the constraints in the specific assessment situation and the validity of the assessment. First, a choice has to be made between whether the test takers revise their own text or a given text produced by another writer. Although the revision of one's own text would be the most realistic representation of revising during the process of writing, revising another writer's text is also a realistic communicative setting. Moreover, several problems arise when measures of revision ability are based on pupils' corrections of their own texts. First, studies have shown that pupils execute only a limited amount of revisions in their own text spontaneously (Van Gelderen, 1997), that skilled revisers postpone revision activities towards the end of the writing process (Van der Hoeven, 1997), and that more changes are made in reviewing another writer's text (Bartlett, 1982). The latter finding may also be valid for adult writers, since every writer will be familiar with the phenomenon of being blind to one's own faults.

Moreover, the change in perspective from writer to reader is cognitively demanding for younger writers (Kellogg, 2008). In addition, it is difficult to perform an independent assessment of a pupil's ability to revise using a text produced by that pupil. Pupils who do not produce lengthy texts and/or errors have a limited chance of displaying their revision skills. In other words, the measure of revision ability partly depends on the amount and quality of the text produced. This problem can be overcome by providing each pupil with the same text that contains the same flaws, thereby giving each pupil the equal chance to display their ability to revise. Lastly, it is difficult to separate the process of text production (i.e.,

generating and formulating) from the process of text revision (i.e., reading and editing the text) because a writer constantly monitors and evaluates his or her output, which means that revisions can take place *during* the production of a text. Given the above, the revision tests evaluated in this study require the test takers to revise a text that was (supposedly) written by a peer.

In adopting a test format in which the test takers are presented with a text, a second consideration is whether or not to indicate the defects that require correction. If these defects are not indicated, the test takers can demonstrate their ability to *detect* flaws as well as their ability to *correct* them. Although this would arguably be the most realistic representation of a real-world language situation, thus enhancing validity (Bachman & Palmer, 1996), scoring difficulties might threaten reliability. That is, scoring rules are needed in order to handle situations in which the test taker improves the text at a point that was not specified as an error, or situations in which the test taker rightly identifies a defect, but subsequently does not change it correctly. Thus, it has to be evident to the test taker which types of corrections are expected and which are not because even a text without actual "errors" can be altered and thus improved at many points. In this respect, the (inter-rater) reliability and validity of a revision test is likely to benefit from a format that indicates the defects that require correction. Therefore, the text format adopted in this study uses another writer's text in which the sentences containing errors are indicated.

### 3.2.4 Response formats

After the text format is chosen, the second decision concerns the response format. When constructing a revision test in which the defects are indicated, the manner in which the responses are corrected by the test taker needs to be specified. In general, two types of answer formats are distinguished, which differ in the amount of input required from the candidate. In a *selected response* format (e.g., multiple-choice, true/false and matching) the test taker chooses a response from a set of given options. A *constructed-response* format (e.g., sentence completion, short answer and essay) on the other hand, requires the test taker to construct the response actively (McMillan, 2008). Response formats vary along a continuum of response construction, in which a multiple-choice format represents the lowest degree of response construction, and a portfolio-performance assessment represents the highest degree (Bennett, 1993; Snow, 1993). The type of constructed responses applied in a revision test (e.g., deleting, adding, or changing words), falls between these two extremes on the scale of response types.

In large-scale assessments, a selected response format in the form of multiple-choice (MC) items is usually favoured over constructed-response (CR) items (Rodriguez, 2003). Several properties of the MC test format account for its popularity. First because it is clear beforehand which of the given response options is the correct answer, MC items can be scored objectively. Objective scoring is straightforward and time efficient, which means that more items can be scored in a given amount of time and for a given amount of money. Thus,

without the need for extra scoring resources, a larger number of items can be admi-nistered, thus increasing test reliability. Furthermore, a large set of test items can increase content validity by enabling a broader coverage of the domain (In'nami &Koizumi, 2009). However, it has been argued that MC items fail to elicit behaviour that represents higher levels of cognitive processing (Campbell, 1999), which would impair construct validity.

Constructed-response items, on the other hand, are gaining in popularity because of the belief that they allow a more direct measurement of certain cognitive processes (Rodriguez, 2003), which arguably leads to a more direct assessment of the trait assessed, subsequently adding to its validity. Furthermore, CR items can also be scored objectively, as long as the items are constructed such that the correct response(s) can be unambiguously stated beforehand (McMillan, 2008). Lastly, with the use of CR items, the probability of guessing is reduced.

MC items, therefore, are considered to offer a greater breadth of domain coverage, whereas CR items offer a greater depth of information on the processes concerned with the trait assessed (Messick, 1993). In other words, both response formats have characteristics that are valuable for the assessment's validity. In past decades, several comparative studies have focused on the differences between the MC and the CR test formats. These studies aimed to determine the extent to which the psychometric characteristics of a test and the trait it purports to assess are influenced by the response format, under the assumption that different formats induce different cognitive processes (Hohesinn & Kubinger, 2011).

In one of these comparative studies, Akerman and Smith (1988) applied the cognitive model in Flower and Hayes (1981) to describe the differences between MC and CR formats. They argued that answering an MC item only requires the process of *reviewing* the answer options, so no *planning* or *translating* is executed. In CR items, the additional process of *generating* a response is required. Although high correlations between stem-equivalent items in both MC and the CR formats are typically found in comparative studies (Rodriguez, 2003), evidence of trait equivalence is limited and unsound, which raises the question of whether MC and CR items measure the same construct (Rodriguez & Traub, 1993). Moreover, Traub (1993) stated that in some domains, including writing, MC and CR response formats seem to measure different constructs. Lissitz and Hou (2012) argued that different response formats do not necessarily elicit the same expression of skills and that CR items may assess skills in a different manner than MC items do. Lastly, MC items were generally found easier then CR items (e.g., In'nami & Koizumi, 2009). In the present study, the validity of both multiple-choice and constructed response revision items is evaluated, as well as the differences in test characteristics of both formats.

### 3.2.5 Test formats

Combining the text formats and response formats described in the preceding sections results in four different formats to test revision ability. Table 2 presents an overview of these test formats; the two test formats under consideration in the present study are indicated in bold.

In the first format, a text containing defects of different categories is presented and accompanied by a series of multiple-choice questions (test format B in Table 2). By answering these questions, test takers choose which action to undertake in order to correct the text. These tests can be scored objectively and efficiently, while at the same time a broad scope of revision activities can be operationalized (cf., In'nami & Koizumi, 2009). However, since test takers are only required to read the text and subsequently choose between several given answers, this format relies heavily on reading comprehension skills, but requires no demonstration of (active) formulating skills. Moreover, because the format is a text with a series of accompanying questions, these MC tests resemble reading comprehension tests at first sight. Hence, both the face validity and the construct validity of the MC test format can be considered impaired.

Table 2. *Array of Test Formats for Revision Tests*

| **Text format I (writer)** | **Text format II (defects)** | **Response format** | **Test format** | |
|---|---|---|---|---|
| Writer's own text | Defects to be detected by test taker | Constructed-response (CR) | Test taker writes a text and subsequently detects and revises defects | A |
| **Other writer's text** | **Defects are indicated** | **Multiple-choice (MC)** | **Test taker chooses the right correction for defects in given text** | **B** |
| | | **Constructed-response** | **Test taker formulates a correction for defects in given text** | **C** |
| | Defects to be detected by test taker | Constructed-response | Test taker detects and revises defects in given text | D |

*Note.* Test formats in **bold** represent the two types of formats under consideration in the present study.

In contrast, constructed response test formats have high (face) validity given that the setting in which test takers actively correct flaws is authentic (see formats A, C and D in Table 2). However, it is difficult to operationalize higher order revision activities, such as adaptation to audience, goal, and text structure in these formats, which may lead to incomplete coverage of the test domain and thus to impaired (content) validity. In addition, the fact that pupils will likely execute the task in various ways can lead to subjectivity in scoring the responses, thus lowering the reliability. The present study attempts to overcome the latter disadvantage by piloting a test format in which the sentences containing flaws are indicated, and the types of permissible corrections are restricted (format C in Table 2).

In the following sections, two types of test formats are evaluated, in which defects in another writer's text are indicated. In section 3.3, the validity of an existing revision test in a multiple-choice format is discussed (test format B in Table 2). Section 3.4 discusses the validity of a piloted constructed-response version of this test (test format C in Table 2).

## 3.3 The validity of a multiple choice revision test

### 3.3.1 Research questions and hypotheses

The present study aims to determine how a revision test can be validly and reliably incorporated into a large-scale assessment of writing ability in primary education. This section focuses on the evaluation of an existing standardised multiple-choice (MC) revision test format that is administered in most Dutch primary schools. To evaluate the validity of the MC revision test, aspects of content and construct validity were considered. The following research questions were addressed:

1. *To what extent does the content of the current multiple-choice revision test represent the content domain of revision ability?*
2. *To what extent do correlations with related constructs (reading, vocabulary, and writing) and non-related constructs (arithmetic) support the construct validity of the current multiple-choice revision test?*

First, the evaluation of content validity focuses on the extent to which the items within the revision test represent the domain of revision ability for novice writers. Given the test format, in which a text is given and defects are indicated (cf. Table 2), certain elements of revision ability are difficult to assess. These elements mainly concern revisions on the conceptual level, such as altering the organisation or coherence of the given text (cf. Figure 2). Hence, the content of the MC revision test is expected to cover only part of the content domain for revision ability.

Second, to evaluate construct validity, correlations between the scores on the MC revision test and the scores on the assessments of related and unrelated constructs are computed. These correlations serve as measures of convergent and divergent validity, respectively. Because correcting a text requires careful reading and the application of linguistic knowledge, writers who are equipped with a broad vocabulary and who are skilled in reading and writing will most likely prove to be better revisers. Scores on the MC revision test are thus expected to show a higher correlation with test scores in related constructs (i.e., reading, vocabulary, and writing), compared to test scores in a non-related construct (i.e., arithmetic).

### 3.3.2 Method

Materials

*Multiple-choice test*

The Cito Entrance Tests for grades 3 to 5 (Entreetoetsen groep 5-7, Cito, 1998-2010) were used to evaluate the MC revision test and provide test scores on related and non-related constructs. The test format adopted in this standardised test is the same format that is used in the Cito End Test for Primary Education (Eindtoets Basisonderwijs), which is administered to pupils in the final year at over 6,000 Dutch primary schools each year. The

Entrance Tests are administered over the course of three days and consist of around 400 multiple-choice items that cover arithmetic, study skills, and several language subjects, as well as 50 items about revision ability. In these items, pupils are given a series of texts (supposedly) written by peers and are prompted to revise these texts by answering multiple-choice questions. Ability scores based on all 50 items were used in this study. Appendix A provides an example of a text with accompanying test items. All answers on the multiple-choice items are recorded on optically readable answer sheets, which are then scored automatically.

*Writing assignments*

Three different essay tasks were selected from the pool of tasks in the Dutch national assessment (Krom et al., 2004). Together, these tasks covered a broad range of communicative goals and text genres (Table 3).

Table 3. *Essay Tasks*

| Task | Description | Communicative goal | Text genre |
|---|---|---|---|
| **A Tigors & Giraks** | Finishing an adventurous story on two tribes | Narrative | Story |
| **B Pookie** | Writing a note describing a lost cat and requesting help | Descriptive / Directive | Leaflet |
| **C Yummie** | Writing a letter to convince a company to accept an incomplete stamp card | Argumentative / Persuasive | Letter |

Participants

Five Dutch primary schools representing different regions, school sizes, and denominations volunteered to participate in the study. In all schools, the Entrance Tests were administered in grades 3 to 5. Additionally, essay assignments were given to a total of 620 pupils, aged 8 to 12 (Table 4). In two schools, participating pupils executed all essay assignments, producing three essays in total. In the other schools, the pupils produced a set of two essays. To avoid sequence effects, one of six possible combinations of tasks was assigned randomly to the pupils. In total, 1,475 essays were collected.

Table 4. *Collection of Essays*

| Number of pupils | Grade 3<br>Age 8/9 | Grade 4<br>Age 9/10 | Grade 5<br>Age 10/11 | Grade 6<br>Age 11/12 | Tasks per pupil |
|---|---|---|---|---|---|
| School 1 | 35 | 48 | 26 | 26 | 3 |
| School 2 | 25 | 25 | 25 | 25 | 3 |
| School 3 | 41 | 31 | 33 | 38 | 2 |
| School 4 | 48 | 34 | 44 | 24 | 2 |
| School 5 | 23 | 17 | 36 | 16 | 2 |
| **Total** | **172** | **155** | **164** | **129** | **1475 essays** |

Analyses

As previously mentioned (cf. Section 3.2.1), multiple tasks rated by multiple readers are needed to assess writing reliably and validly. Following Coffman's calculation of a validity coefficient (Coffman, 1966), Van Schooten and De Glopper (1990) stated that a minimum of three writing tasks rated by a minimum of two assessors or two writing tasks rated by three assessors is needed in using essays as a valid measure of construct validity (Van Schooten & De Glopper, 1990, p. 98). These numbers broadly correspond to findings in more recent research (Schoonen, 2005; Van den Bergh, et al., 2012). Within the present study, all essays were evaluated by two raters. A rating scale with anchor essays was used to evaluate the writing products. Raters were given a list of analytical questions accompanied by a rating scale with exemplar essays per aspect (content, structure and correctness). In the final question per aspect, they were asked to provide an overall score (A to G) for this particular aspect by placing the essay on the provided scale of writing ability (Figure 3). Satisfactory inter-rater reliability scores for different tasks and aspects were found, ranging between .75 and .86 (cf. Chapter 2).

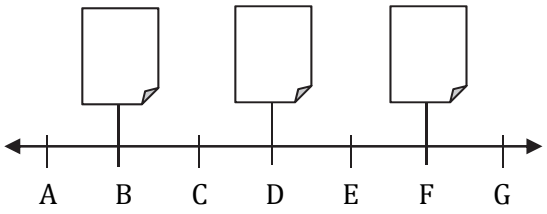



*Figure 3*. Scoring options for the overall score.

### 3.3.3 Results

Content validity

To evaluate the extent to which the content of the MC revision test was a valid representation of revision ability, the content of the test was matched with the elements of revision ability listed by Hayes et al. (1987). Figure 4 shows the item categories specified in the MC revision test on the left (A). A list of revision activities is given on the left. The dotted lines indicates matches between the test content (A) and elements of revision ability (B). As Figure 4 shows, not all elements of revision ability are represented in the revision test. At the conceptual level, no items on text structure, coherence and organization are included (cf. Figure 4, elements in *italics*). At the textual level, the revision of spelling faults is not incorporated.

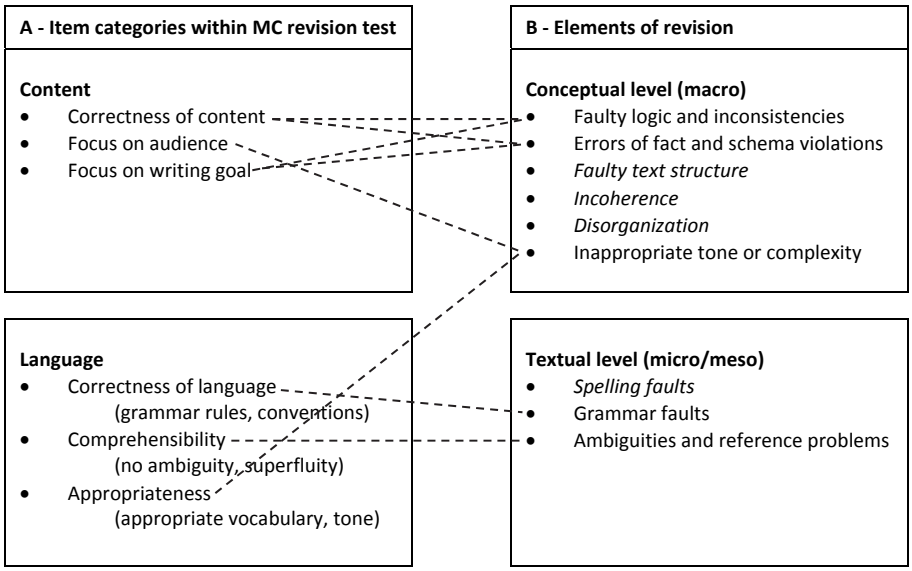



*Figure 4.* Matching the content of the revision test to revision activities.
(Hayes et al., 1987)

Table 5 shows the number of test items per category, further indicating that the focus was on the assessment of formulating skills, not on the conceptual level of text revision. Less than one third of the test items concerned conceptual changes, such as correcting inconsistencies in content and inappropriate tone or complexity.

Table 5. *Distribution of Test Items per Category*

| **Item categories within MC revision test** | **Text level** | **Number of items (%)** |
|---|---|---|
| **Content** | | |
| • Correctness of content | macro | xx (4%) |
| • Focus on audience | macro | xxx (6%) |
| • Focus on writing goal | macro | xxxx (8%) |
| **Language** | | |
| • Correctness of language | micro/meso | xxxxx xxxxx xxx (26%) |
| • Comprehensibility | micro/meso | xxxxx xxxxx xxxxx xxxxx xxx (46%) |
| • Appropriateness | macro | xxxxx (10%) |

Construct validity

Table 6 shows the Pearson correlations between the scores on text revision and the scores on three other sections of the Entrance Test. The results showed high correlations between text revision and related (language) constructs in the Entrance Test and significantly lower correlations between text revision and arithmetic.

Table 7 through Table 9 report the correlations between the results on the Entrance Test and the scores on the three different writing assignments A to C (cf. Table 3). These results showed that correlations between text revision and other language subjects within the Entrance Test (vocabulary and reading) were significantly higher than the correlations

between text revision and all three writing tasks. Correlations between text revision and arithmetic were lower compared to the language subjects, but they were significantly higher than the correlations with the writing tasks. A similar pattern was found in comparing correlations between vocabulary and reading to arithmetic and the writing tasks.

Table 6. *Correlations between Several Test Subjects in the Entrance Test (grades 3 to 5)*

| **N = 563** | **Text revision** | **Vocabulary** | **Reading** | **Arithmetic** |
|---|---|---|---|---|
| Text revision | 1 | .78*** | .84*** | .69 |
| Vocabulary | | 1 | .75*** | .64 |
| Reading | | | 1 | .65 |
| Arithmetic | | | | 1 |

*** p= 0,000 (given correlation vs. correlation with arithmetic)

Table 7. *Correlations between Subjects in the Entrance Test and Writing Assignment A*

| **N = 273** | **Text revision** | **Vocabulary** | **Reading** | **Arithmetic** | **Writing task A** |
|---|---|---|---|---|---|
| Text revision | 1 | .79*** | .85*** | .69* | .59 |
| Vocabulary | | 1 | .74*** | .63 | .57 |
| Reading | | | 1 | .67* | .57 |
| Arithmetic | | | | 1 | .45 |
| Writing task A | | | | | 1 |

*** p ≤ 0,001; * p < 0,05 (given correlation vs. correlation with writing task A)

Table 8. *Correlations between Subjects in the Entrance Test and Writing Assignment B*

| **N = 269** | **Text revision** | **Vocabulary** | **Reading** | **Arithmetic** | **Writing task B** |
|---|---|---|---|---|---|
| Text revision | 1 | .81*** | .83*** | .66** | .52 |
| Vocabulary | | 1 | .75*** | .63** | .47 |
| Reading | | | 1 | .64 | .45 |
| Arithmetic | | | | 1 | .30 |
| Writing task B | | | | | 1 |

*** p ≤ 0,001; **p<0,01 (given correlation vs. correlation with writing task B)

Table 9. *Correlations between Subjects in the Entrance Test and Writing Assignment C*

| **N = 250** | **Text revision** | **Vocabulary** | **Reading** | **Arithmetic** | **Writing task C** |
|---|---|---|---|---|---|
| Text revision | 1 | .78*** | .84*** | .72*** | .53 |
| Vocabulary | | 1 | .76*** | .66*** | .49 |
| Reading | | | 1 | .66** | .51 |
| Arithmetic | | | | 1 | .39 |
| Writing task C | | | | | 1 |

*** p ≤ 0,001; **p<0,01 (given correlation vs. correlation with writing task C)

### 3.3.4 Discussion

The present study evaluated the validity of a multiple-choice (MC) revision test. First, the extent to which the test items covered the content domain of revision ability was determined. Figure 4 shows that the MC test offered a diverse array of test items and covered a fair amount of the content domain for text revision. The only two revision activities that were not incorporated in the MC revision test were text structuring and spelling; the latter is assessed as a separate skill in the Cito Entrance Tests. The current MC format does not allow the sensible evaluation of text (re)structuring skills because they would have to be measured indirectly by asking the pupil to relocate a certain part of a text by indicating where to place it. In addition, the use of a text with a faulty structure possibly influences the other items accompanying that text. It is possible that a digital test format would allow a more direct assessment of text structuring skills by asking the pupil to drag text parts to their destined location, thereby allowing them to review the results of their action.

Further analysis of the test content showed a focus on revisions concerning the textual level, as opposed to the conceptual level (cf. Table 5). This finding, however, does not necessarily impair the validity of this assessment. First, the revision activities of novice writers take place mainly in the translation stage within the writing process (Kellogg, 2008), which leads to changes on a textual level instead of a conceptual level (cf. Figure 2). Second, the validity of a test depends on the intended use of the test scores (Kane, 2006). Hence, the underrepresentation of revision activities at the conceptual level does not impair the validity if the MC revision test is aimed at assessing pupils' ability to revise a text at the textual level. In this study, the fact that the revision test is incorporated within a large-scale writing assessment further supported validity because the conceptual level of text production was evaluated elsewhere in the assessment.

As expected, scores on the MC revision test showed a lower correlation with the scores on arithmetic, compared to other language constructs, indicating that the ability assessed by the revision test is indeed closely related to language constructs that are believed to be interrelated with writing ability (i.e., vocabulary and reading) instead of non-language construct arithmetic (Table 6). Hence, these correlations are evidence of divergent and convergent validity. However, correlations between scores on text revision and scores on writing tasks did not exceed correlations between other subjects in the Entrance Test, including arithmetic (Table 7 through Table 9). This result was unexpected because previous research reported high correlations between scores on revision tests and writing tasks (cf. Table 1).

A possible explanation for the low correlations between the revision test and the writing tasks could be that the revision tests were administered at a different time and in a different setting than the writing assignments were. Revision scores were taken from the Cito Entrance Test, which is considered a fairly high-stakes test that is administered during the course of three mornings and thus resembles the setting of the mandatory test taken at the end of primary education. The writing assessments, on the other hand, were

administered between classes and resembled a (low-stakes) classroom assessment. These different circumstances may well have caused the variance in the two performances by the same pupil. An additional explanation could be that the evaluation of writing tasks focused mainly on the conceptual aspects of a writing product, such as generated content, focus on writing goal and overall text structure. Revision tests, on the other hand, concerned mainly "low level" skills (Camp, 2009), such as grammar and vocabulary, a stance that is supported in this study (Figure 4, Table 5).

With respect to the test format, multiple-choice questions are regarded an indirect and inauthentic assessment of an active skill (Rodriguez, 2003). Hence, the validity of a test in which editing a text is assessed by answering multiple-choice questions is considered impaired. Furthermore, this kind of MC revision test could be easily mistaken for a test on reading comprehension, thus providing further evidence of impaired (face) validity.

The evaluated MC revision test mainly covers revision ability at the textual level. This focus on formulating skills at the sentence and word level is suited for an assessment in primary education because most novice writers edit their texts at the surface level (Kellogg, 2008). However, because answering multiple-choice questions does not require the actual activity of editing a text, face validity was seriously impaired. It is possible that this drawback could be overcome by creating a constructed-response (CR) version of the current revision test. Because CR assessments are considered a more authentic representation of the ability assessed, validity would be supported (Bachman & Palmer, 1996). The following section reports a pilot-study in which a newly developed CR revision test was trialled.

## 3.4 The validity of a constructed-response revision test

### 3.4.1 Research questions and hypotheses

The aim of this study is to investigate how to validly assess revision ability within a large-scale assessment of writing in primary education. For this purpose, the validity of a multiple-choice (MC) revision test was evaluated (Section 3.4). Although satisfactory results were found for construct and content validity, the MC format arguably lacks authenticity and thus face validity. Therefore, this section discusses the pilot test of a newly developed constructed-response (CR) version of the MC revision test and compares it to an MC version of the same test.

In order to evaluate the validity of the CR revision tests, the extent to which the items within the revision test represent the domain of revision ability for novice writers was investigated. The following research questions were addressed:

1. *To what extent does the content of the piloted constructed-response revision test represent the content domain of revision ability?*
2. *To what extent does the piloted constructed-response revision test provide for equivalent test characteristics (i.e., reliability, difficulty, and discriminative power) when compared to the multiple-choice version?*

Because the format of the chosen test is a text in which defects are indicated (cf. Table 2), certain revision activities cannot be assessed. The results presented in Section 3.3 showed that the content of the MC revision test covered only part of the content domain for revision ability. For the CR version of the revision test, a further confined coverage of the content domain is expected because not all multiple-choice items will allow the conversion into constructed-response items. Consequently, fewer test items will be available, which means that some parts of the domain are likely to be un- or under-represented. In general, the CR test format is expected to cover mainly the textual component of revision, not the conceptual component (cf. Figure 2).

With respect to psychometric characteristics, the CR test format is expected to elicit results of comparable reliability. Furthermore, the discriminative power of the test is not likely to be influenced by the test format. On the one hand, less able pupils are expected to have more difficulties in actively correcting the defects than in choosing the correct alternative; on the other hand, highly able pupils might want to change more than is required, thus risking the loss of points. Lastly, a high correlation between the two test formats is expected, indicating that both tests assess the same trait.

### 3.4.2 Method

Participants

This study was conducted as a pilot study within the 2009 Dutch national assessment on writing ability (Kuhlemeier, Van Til, Hemker, De Klijn, & Feenstra, 2013). Eighty schools participated in the assessment, of which >1600 pupils were sampled to participate. Within this test population, 522 pupils enrolled in grade 6 were assigned to participate in the pilot study. An incomplete design was used to assign the participants a combination of two tasks in different test formats: multiple-choice (MC) and/or constructed-response (CR) (Appendix B).

Materials

*Test format*

In the new format of the Cito MC revision test, pupils are instructed to revise the given text actively. A so-called interlinear test format was adopted to facilitate editing (cf. Wesdorp, 1974, Schoonen, 1991). The interlinear test consists of one or more double-spaced texts containing defects. Test takers are asked to correct these defects by using the space between the lines—hence, the term "interlinear." In order to prevent highly divergent responses, specific instructions are given on the type of corrections that are allowed. Appendix C is an example of this CR test format. To enable objective scoring, the types of changes that pupils were allowed to make in order to improve the text were limited. Possible corrections were *deleting*, *adding*, *switching* or *substituting* words or phrases. Pupils were given short instructions that included an illustration of all admissible corrections, as shown in Figure 5.

| | |
|---|---|
| Afgelopen zaterdag ~~toen~~ ging ik naar mijn oma.<br>[Last Saturday ~~then~~ I went to my grandma.] | *deleting* |
| eerst<br>We gingen naar buiten. Daarna maakten we teams.<br>first<br>[We went outside. After that, we made teams.] | *adding* |
| Als ik vrij ben, ik ga graag voetballen.<br>[When I am free, like I to play football.] | *switching* |
| zijn<br>Mijn hobby's ~~waren~~ tekenen, judo en gamen.<br>are<br>[My hobbies ~~were~~ drawing, judo and gaming.] | *substituting* |

*Figure 5.* Admissible corrections within the CR revision test.

*Test versions*

Two short texts were written: text A and text B. Twenty-seven multiple-choice items were constructed, 12 of which pertained to text A and 15 to text B. These items were also the basis of the constructed-response items. Some of the MC questions could not be converted into CR questions because of the limited options for corrections (cf. Figure 5), so the CR version consisted of 15 items in total: 7 items for text A, and 8 items for text B. To compare the test quality of the newly developed test items, a MC revision test of 50 items was used. This test was taken from the Cito Entrance Test for primary school pupils (Entreetoetsen groep 5-7, Cito: 1998-2010).

*Scoring*

The scoring script was formulated by the constructors of the test items. The constructors independently rated a sample of items. They then reached agreement on the scoring rules and constructed a scoring instruction based on a few simple guidelines (Table 10). This instruction was then implemented by an experienced rater who scored all responses on the open-ended versions of the test.

Table 10. *Scoring Rules for CR Revision Test*

| **1 point** | **0 points** |
|---|---|
| target error is corrected | target error is ignored |
| target error is corrected, BUT other error is added | target error is altered, BUT incorrectly |

Analyses

Classical test theory (Novick, 1966) was used to compute the reliability, difficulty and discriminative power of both the existing MC test and the newly developed CR test. The differences in test characteristics between these tests were analysed using an independent samples t-test.

### 3.4.3 Results

*Content validity*

Figure 6 shows the item categories specified for the constructed-response (CR) revision test on the left (A). Revision activities (B) are listed on the right. The dotted lines indicate matches between the test content (A) and the elements of revision ability (B). Hence, Figure 6 shows that only a limited number of elements is represented in the CR revision test. In the conceptual level, only the inappropriateness of tone or complexity is represented, while no items on content, text structure, coherence and organization are included (cf. Figure 6, elements in *italics*). The revision of spelling faults is not incorporated into the textual level.

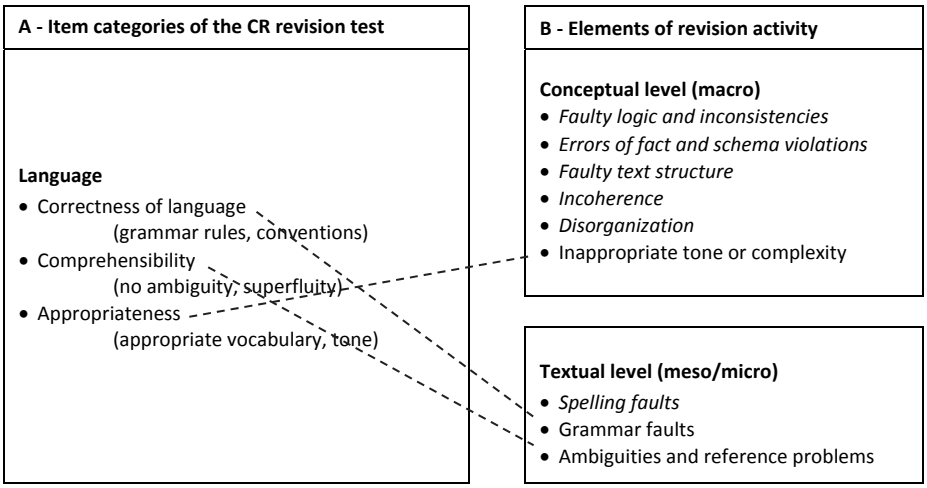



*Figure 6*. Matching the content of the CR revision test to revision activities.
(Hayes et al., 1987)

Table 11 shows the number of test items per category for the MC and CR revision tests. A comparison of both methods showed an even stronger focus on the revision of language elements than on content elements in the CR version. Apparently, not all content of the MC items was suitable for assessment in an open-ended question format. Table 12 presents some examples of items for which conversion into a CR item proved impossible. Appendix D presents the results of the content analyses of both the MC and the CR test formats.

Table 11. *Distribution of Test Items per Category*

| **Item categories of the revision test** | **Text level** | **Number of items (%) CR** | **Number of items (%) MC** |
|---|---|---|---|
| **Content** | | | |
| Correctness and completeness | macro | – (0%) | – (0%) |
| of content | macro | – (0%) | x (4%) |
| Focus on audience | macro | – (0%) | x (4%) |
| Focus on writing goal | | | |
| **Language** | | | |
| Correctness of language | micro/meso | xxxxx xxxxx (67%) | xxxxx xxxxx xxxxx (56%) |
| Comprehensibility | micro/meso | xxxx (27%) | xxxxx xxxx (33%) |
| Appropriateness | macro | x (7%) | x (4%) |

Table 12. *Examples of MC Items for which No CR Equivalent Could Be Constructed*

| | **Item** | **Category** | **Text level** |
|---|---|---|---|
| (1) | Which title should Emma choose for her text? | Correctness and completeness of content | macro |
| (2) | Joan finished her mail with *Doei*! (line 10). Which of the below greetings does not fit the rest of the email? | Focus on audience | macro |
| (3) | Which sentence does not fit the rest of the text? | Focus on writing goal | macro |

Test characteristics

The test characteristics shown in Table 13 indicate that the reliability of both formats of the newly developed test were considered "satisfactory" (alpha .60 ≤ r < .70), according to the standards for assessment at the group level, which were determined by the committee on test evaluation in the Dutch Institute of Psychology (COTAN, 2010). In order to reach the standard "good" (alpha ≥ .70), 40 CR items were needed, compared to 51 MC items (based on reliability prediction using the Spearman Brown prophecy formula). However, this finding does not imply that CR tests are more efficient because CR items are expected to elicit longer response times when compared to MC items.

The average difficulty of the test items is represented by the observed proportion of pupils that answered the items correctly (p-value). On average, the p-values of the CR test items were lower (M=.62, SD=.17) than the MC items (M=.73, SD=.21) were. However, this difference is not significant (t(40)=1.80, p>.05). The average discriminative power is represented by the correlation between the score on the separate test items and the total test score ($R_{it}$). In both test versions, $R_{it}$ values were considered "good" (>.30), according to the standards (COTAN, 2010). On average, $R_{it}$ values for the CR test were higher (M=.43, SD=0.07) than those for the MC items (M=.39, SD=.10). This difference was not significant (t(40)=-1.15, p>.05). Furthermore, a high observed correlation (r=.78) between scores on the two test versions was found.

Table 13. *Test Characteristics for CR and MC Versions*

| | **CR questions (n=15)** | **MC questions (n=27)** |
|---|---|---|
| Test reliability (alpha) | .60 | .68 |
| Av. item difficulty (p-value) | .62 | .74 |
| Av. discriminative power ($R_{it}$) | .43 | .39 |

In order to evaluate the quality of the newly developed multiple-choice items as a criterion for comparison, the MC items were also compared with an existing multiple-choice test of proven quality. Comparison between the test characteristics of the new MC test and the existing version in the Cito Entrance Test showed that p-values for the existing multiple-choice items were higher (M=.72, SD=.13) than the p-values of the new MC test (M=.74, SD=21). This difference was not significant (t(37)=-.353, p>.05). Furthermore, the existing items had higher discriminative power (M=.46, SD=.08), than the newly developed items (M=.39, SD=.10) did, which was a significant difference (t(75)3.33, p<.05).

### 3.4.4 Discussion

Because the multiple-choice format is considered unauthentic in assessing text revision, the present study piloted a constructed-response (CR) version of the revision test. In this test format, sentences containing defects were indicated, and pupils had to correct these flaws by writing between the (double spaced) lines. The (face) validity was improved because that test takers directly edited the given text, compared to the indirect approach in which the test takers respond to multiple-choice questions. The pilot test format also represented an authentic communicative setting. Pupils were given a text (supposedly) written by a peer and were asked to correct it. Therefore, the CR test format is considered to resemble a real-world language situation of revision, which enhances its predictive power and, thus, overall validity (Bachman & Palmer, 1996).

However, the content analysis of the CR test revealed that only a limited part of revision ability could be assessed in this format (Figure 6, Table 11, and Appendix D). Compared to the contents of the MC test format, the results showed an even more apparent focus on formulating skills than on conceptual skills. Most items concerned the micro text level of the correctness of language, and only one item addressed a correction on the macro/meso level.

The psychometric characteristics of both the MC and CR items were compared to evaluate further the quality of the newly developed CR items. As Table 13 shows, no significant differences were found. Similar to earlier findings by Hohensinn and Kubinger (2011) and In'nami and Koizumi (2009), the CR items seemed slightly more difficult. Contrary to expectations, the CR test was slightly more reliable than the MC test. Rodriguez (2003) reported similar findings and attributed this difference to the fact that the answer options in MC items provide clues, so random error is introduced and reliability is subsequently reduced. Furthermore, a high observed correlation (.78) between both versions of the piloted test was found. Given this correlation, both the CR and the MC formats appear to measure the same latent trait to a great extent, but the formats may elicit different skills and/or assess skills differently (cf. Hohensinn & Kubinger, 2011; Lissitz & Hou, 2012).

Comparison of the test characteristics of the existing MC test and the newly developed MC test showed no difference in difficulty between the two tests. Furthermore, even though the discriminative power of the piloted MC test was lower than that of the existing test, it was still considered "good" (i.e. >.30), according to the standards (COTAN, 2010). This result indicates that the newly developed MC test is indeed an appropriate criterion when comparing multiple-choice items to constructed-response items.

## 3.5 Discussion and Conclusion

The present study explored the assessment of revision ability in primary education. For this purpose, two formats of revision tests were evaluated: a multiple-choice (MC) version and a constructed-response (CR) version (cf. Table 2). The content analyses of both formats of the revision tests showed a clear focus on formulating skills. Therefore, these tests cannot be considered a valid representation of all levels of cognitive ability, as shown in Figure 2. Instead, it seems justifiable to claim that the evaluated tests measured the ability of pupils to perform revision activities on the textual level. In this level, the MC format offers a broader domain representation, compared to the CR format. In the MC format, however, the domain was assessed in a more or less indirect manner because answering multiple-choice items does not require the active correction of the indicated faults. The CR test format, on the other hand, offered less breadth in domain representation, but the manner of assessment was more authentic, thus offering improved face validity.

Assessing the revision ability of novice writers

In constructing a revision test, the purpose of the test and its target population determine which elements of revision (cf. Figure 2) should or should not be incorporated. As is shown by Sommers (1980) and Faigley and Witte (1981), even writers enrolled in the first year of university rarely (spontaneously) revise their own text on the conceptual level by making changes in macro structure. Instead, these students execute changes at the surface level, such as rewording and correcting grammatical errors. Furthermore, studies on the cognitive development of writing skills indicated that for inexperienced writers, (i.e., writers in the "knowledge telling phase" (Bereiter & Scardamalia, 1987) the interaction between the writer's thoughts and their representation in generated text is cognitively very demanding (Kellogg, 2008). Hence, for these writers, the act of producing a text is dominated by the need to translate their thoughts on paper, while revision activities are largely defined by checking this translation.

Thus, the ability to make conceptual changes in a given text will be beyond the mastery of most primary school pupils. Hence, the assessment of revision ability should focus on editing on the (textual) surface level instead of on the conceptual level. Furthermore, asking for textual improvements only is likely to suit the novice writer's level of writing ability (Hayes et al., 1987, p. 196). Novice writers are mainly oriented toward their own activities, and their revision activities are limited to adding and deleting, whereas expert writers are reader-oriented and adjust and reshape their text in order to achieve their intended writing goal (Breetveld, 1991; Kellogg, 2008; Van der Pool, 1995). Moreover, the ability to perform these conceptual skills is covered by writing assignments in which novice writers do not necessarily need to switch from the generating role to an evaluating one. Research on the cognitive processes involved in writing showed that novice writers mainly focus on "getting language onto paper," leaving no capacity to consider higher order processes (Bereiter, 1980).

Although revising a text as a reader instead of a writer may appear to be an indirect approach, being able to revise another writer's text is a valuable skill. By revising, pupils show their awareness of several criteria for successful communication (Van Gelderen, 1997). Bartlett (1982) found that detecting reference problems between sentences is easier from the viewpoint of a reader than that of an author. Furthermore, in a study by Chanquoy (2001), the frequency of revisions was found to be higher when the revision is postponed to the end of the writing process.

Research has also shown that novice writers do possess the skills of revising, but need to be triggered in order to show their potential (Bereiter & Scardamalia, 1987). A revision test allows novice writers to showcase their ability to edit a text without being hindered by higher order processes they do not yet fully command. Hence, the revision assignment allows test developers to evaluate specific linguistic issues independently of the text produced by the pupil. On the one hand, this enables pupils who have trouble executing other components of the writing process (i.e., planning and/or generating) and are therefore less able to produce text to demonstrate their ability in revision. On the other hand, it provides researchers and/or teachers with detailed information on the language ability of pupils. In addition, tests on specific components of writing, such as revision tests, can be employed to comprehend further the processes involved in writing (Schoonen, 2011; Cushing Weigle, 2002). Similarly, Dean and Quinlan (2010) advocated considering both the *product* and the *process* of writing ability because the final drafts of essays do not necessarily reveal the process of text production. Gaining insight into revision ability will help to further the understanding if the linguistic skills needed to produce a high quality text, thus providing useful information to both educators and test developers.

Constructed response versus multiple choice

With regard to the response format of a revision test, a valid assessment should represent the act of revising a text as closely as possible. The results of the present study indicated that a constructed response format in which test takers are asked to improve the text by editing indicated sentences allows for the reliable direct assessment of revision ability with high (face) validity. However, to enable objective scoring, domain coverage is limited to certain aspects at the textual level. In assessing revision ability in a national assessment of writing in primary education, it would be justifiable to focus on formulating skills only because conceptual aspects, such as structure, content and organization, are directly assessed by means of a writing assignment. In addition, a task that focuses on formulating skills would be a useful addition to a writing assessment because students are given the chance to avoid formulating problems when writing an essay. Instead of assessing formulating skills by offering specific tasks, revision tests enable the assessment of specific formulating issues in a realistic communicative setting.

The evaluated multiple-choice format, on the other hand, is considered less representative of the cognitive processes involved in revising a text, but it offers a broader

domain coverage. If it is desirable within a given assessment context, this test format could even allow for the assessment of the conceptual level of text revision, which would further broaden the domain coverage. Because the scores on both the multiple-choice test and the constructed response test were highly correlated, it appears that both test formats measured the same trait. Hence, a combination of formats might promote the assessment of revision ability by offering both breadth of domain coverage and depth of cognitive process representation.

Future research in the area of text revision could focus on further specifying the construct of revision ability, as well as cognitive processes and development involved in learning to be a reader while writing. These insights could then be employed to support language teachers in creating suitable teaching methods and test developers in creating sound assessments of revision ability. In addition, a digital test format may allow the objective assessment of a wider range of revision activities, including editing at the conceptual level. Given the fact that revision activities of novice writers mainly take place at the textual level, test items aimed at the conceptual level are likely to identify highly skilled pupils that are able to execute revisions beyond the surface level.

**Conclusion**

The results of validity and test characteristics reported in the present study showed that a constructed response test format is a valid and reliable method for the assessment of the revision ability of novice writers. A constructed-response task offers an authentic representation of the process of editing a text at the surface level. In addition, test reliability and discriminative power were found satisfactory. However, the domain coverage of (paper-based) constructed response revision tasks is limited because not all elements of the revision process are suitable for assessment in an open-ended format, due to scoring difficulties. A multiple choice format, on the other hand, is less authentic, but it offers a larger array of test questions and hence broader domain coverage. Consequently, the construction of a revision test that combines both item formats is recommended to improve the assessment of revision ability.

## Appendix A – Multiple-choice format for revision test

*Ik ben een jongen van 12 jaar en ik wil graag schrijven met iemand met dezelfde hobby's. Mijn hobby's zijn schaatsen, tekenen en ik houd ook van popmuziek. Als je zin hebt, schrijf dan wat je vindt van deze hobby's aan:*
*Paul van Veen*
*Postweg 12*
*2601 PS Arnewoude*

Lisette las deze advertentie in een jeugdblad. *Ze* schreef de volgende brief aan Paul.

1 Nieuwersloot, 26-1-2008

2 Hoi Paul,

3 Ik ben een meisje van 12 jaar en ik heb geen
4 broertjes en zusjes. Het lijkt me leuk om met je
5 te schrijven want ik heb bijna dezelfde hobby's
6 als jij. Ik heb wel een hele lieve hond. Die heet
7 Basja. Om te beginnen ben ik gek op schaatsen.
8 Ik rijd op een kunstijsbaan en maak toertochten.
9 We wonen op een kwartier afstand van de
10 ijsbaan af en achter ons huis kunnen we zo op
11 de schaats stappen om een tocht te maken,
12 tenminste als er voldoende ijs ligt. Deze winter
13 heb ik er al heel wat gemaakt.

14 Mijn andere hobby's zijn tekenen en ook lezen.
15 Ik teken graag strips. Die bedenk ik zelf. Ze
16 gaan over van alles en nog wat. Laatst heb ik
17 er een over een toertocht gemaakt. Die doe ik
18 bij deze brief. Wat teken jij het liefst? Als jij ook
19 graag strips tekent, kunnen we misschien een
20 soort vervolgverhaal in elkaar zetten. Dan
21 tekenen we ook iedere keer een strip. We
22 sturen dat naar elkaar op. Dat lijkt me leuk. Zo,
23 nu houd ik op met schrijven, omdat het
24 namelijk zo is, dat ik over ongeveer een half
25 uur de deur uit moet, omdat ik naar de
26 schaatstraining moet. Stuur je me ook gauw
27 een brief terug?

28 Groeten van Lisette

29 Lisette Koch
30 Langedijk 15a
31 8877 JE Nieuwersloot

**Opgave 7**
*Ik heb wel een hele lieve hond. Die heet Basja.*
(r. 6 en 7)
Wat kun je het best met deze zinnen doen?

A Zo laten staan.
B Plaatsen voor: Het ... (r. 4)
C Plaatsen voor: Deze ... (r. 12)
D Plaatsen voor: Mijn ... (r. 14)

**Opgave 8**
*We... ligt.* (r. 9 t/m 12)
Welk woord is in deze zin overbodig en *moet* worden weggelaten?

A af
B zo
C tenminste
D voldoende

**Opgave 9**
Wat had Lisette beter kunnen schrijven in plaats van: ... *er al heel wat* ... (r. 13)?

A al heel wat rondjes
B al heel wat strips
C al heel wat tekeningen
D al heel wat toertochten

**Opgave 10**
*Zo, nu houd ik op met schrijven, omdat het namelijk zo is, dat ik over ongeveer een half uur de deur uit moet, omdat ik naar de schaatstraining moet.* (r. 22 t/m 26)
Wat kun je het best met deze zin doen?

A Zo laten staan.
B Vervangen door: Zo, nu houd ik op met schrijven. Ik moet namelijk over ongeveer een half uur de deur uit, omdat het namelijk zo is dat ik naar de schaatstraining moet.
C Vervangen door: Zo, nu houd ik op met schrijven. Ik moet namelijk over ongeveer een half uur de deur uit, omdat ik naar de schaatstraining moet.
D Vervangen door: Zo, nu houd ik op met schrijven, omdat ik namelijk over een half uur de deur uit moet, omdat het namelijk zo is dat ik naar de schaatstraining moet.

## Appendix B – Test design for pilot study

| | Existing items | | | | Newly developed items | | | |
|---|---|---|---|---|---|---|---|---|
| **Tasks** | **MC 1** | **MC 2** | **MC 3** | **MC 4** | **MC 5** | **MC 6** | **CR 1** | **CR 2** |
| **N of items** / **Booklet** | 12 | 13 | 12 | 13 | 10 | 15 | 7 | 8 |
| 1 | 1 | 2 | | | | | | |
| 2 | | 1 | 2 | | | | | |
| 3 | | | 1 | 2 | | | | |
| 4 | | | | 1 | 2 | | | |
| 5 | | | | | 1 | 2 | | |
| 6 | | | | | | 1 | 2 | |
| 7 | | | | | | | 1 | 2 |
| 8 | 2 | | | | | | | 1 |
| 9 | | | | | 2 | 1 | | |
| 10 | | | | | | 2 | 1 | |
| 11 | | | | | | | 2 | 1 |
| 12 | | | | | 1 | | | 2 |

## Appendix C – Constructed-response format for revision test

Aan: Dolle Dwaze Dieren
Onderwerp: Doorgedraaid!
Bijlage: Doerak draait door.avi

Hoi Dolle Dwaze Dieren!
Jullie hebben een keigoed programma met superleuke dieren. (1) Nou past mij hond Doerak helemaal in jullie programma. (2) Hij is het allergekste hondje die je maar kunt vinden. (3) Als op de eerste maandag van de maand de sirenes gingen, zit Doerak volop mee te janken. (4) Ze doet zijn kop omhoog, zijn ogen dicht en jankt. Wat een herrie man! (5) En als hij zijn staart ziet, hij wil hem pakken en rent hij rondjes. Net zolang tot hij het puntje van zijn staart in zijn bek heeft.
Maar het gekste is nog wel: (6) als ik 's ochtends mijn fiets pak om naar school te gaan, sprong hij achterop! Als dat geen filmpje waard is! Daarom stuur ik jullie hier een filmpje voor jullie rubriek: Doorgedraaid! (7) Ik hoop dat jullie het willen uitzenden in uw programma.

Doei!
Johan Verwegen
johan99@hotmail.com

## Appendix D – Content analyses of constructed-response revision test

| Content | | | Language | | | | | |
|---|---|---|---|---|---|---|---|---|
| Correctness and completeness of content | Focus on audience | Focus on writing goal | Correctness | Comprehensibility | Appropriateness | **Item description** | **Text level** | **MC version** | **CR version** |
| | | | | | x | manner of addressing reader | macro | x | x |
| | | | x | | | personal pronoun | micro | x | x |
| | | | x | | | demonstrative pronoun | micro | x | x |
| | | | x | | | word order subordinate clause | micro | x | x |
| | | | | x | | order of events | meso | x | |
| | | | x | | | verb tense | micro | x | |
| | | | | x | | synonym | micro | x | |
| | | | x | | | gender | micro | x | x |
| | | | | x | | synonym | micro | x | |
| | | | x | | | verb tense | micro | x | x |
| | | | x | | | word order | micro | x | |
| | x | | | | | appropriate greeting | macro | x | |
| | | x | | | | appropriate title | macro | x | |
| | | | | x | | redundant word | micro | x | x |
| | | | | x | | superfluous information | macro | x | x |
| | | | x | | | word order subordinate clause | micro | x | x |
| | | | | x | | order of events | meso | x | |
| | | | x | | | verb tense | micro | x | x |
| | | | x | | | verb tense | micro | x | x |
| | | | x | | | preposition | micro | x | |
| | | | | x | | connective | meso | x | x |
| | | | x | | | word order subordinate clause | micro | x | x |
| | | | | x | | paragraph ending | meso | x | |
| | | | | x | | conjunction | meso | x | |
| | | | x | | | preposition | micro | x | x |
| | | | x | | | structure | meso | x | |
| | | | x | | | verb tense | micro | | x |

# 4 Measuring text complexity: Exploring the usability of automated essay evaluation for novice writers



## 4.1 Introduction

Recent developments in natural language processing (NLP) techniques have enabled the automated analysis of a large array of textual features. Using machine learning techniques, applications have been developed to perform automatically part-of-speech tagging, morphological and semantic analyses, and syntactical parsing. Hence, linguistic information is extracted from texts, providing information about word usage, sentence construction and textual coherence. These features can be combined into measures of text complexity, which in turn can serve to achieve evaluative goals, such as readability prediction and the assessment of text quality. The present chapter explores the use of text complexity measures as indicators of text by novice writers.

### 4.1.1 Assessing text quality

In writing assessment, text quality is commonly interpreted as a measure of writing ability, under the assumption that skilled writers will produce high quality texts (Deane & Quinlan, 2010; McNamara, Crossley, & McCarthy, 2010; Witte & Faigley, 1981). To evaluate text quality, writing tasks are usually deployed in order to collect writing products that are comparable across test takers. The quality of these written products is typically evaluated by one or more trained raters who are guided by scoring instructions. Text quality, however, is notoriously difficult to evaluate reliably and validly, as shown in numerous studies over time (Breland, Camp, Jones, Morris, & Rock, 1987; Godshalk, Swineford, & Coffman, 1966; Cushing Weigle, 2002; Knoch, 2011).

First, consistent scoring of text quality is problematic because of the effects of rater subjectivity on the assigned scores (cf. Meuffels, 1994). Scores assigned by different raters diverge because of variations in emphasis, severity or rating circumstances. Similar rater effects pose a threat to the consistency of scores given by a single rater over time. Furthermore, undesirable effects of both the specific writing tasks assigned and the aspects of writing quality that are scored have been found (Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). Nonetheless, essay tasks enhance the validity of a writing assessment because of their close resemblance to different real-world language situations (Bachman & Palmer, 1996). Despite the aforementioned scoring difficulties, test tasks that involve the production of written text are therefore considered essential when test writing ability is targeted (Blood, 2012).

To compensate for the effects of both raters and tasks, several tasks rated by multiple raters are necessary in order to evaluate text quality reliably (Schoonen, 2005, Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). Automated essay scoring (AES) can assist in overcoming these scoring problems by applying NLP techniques objectively to extract text features that are indicative of text quality. These features can be used to derive scores that offer the additional benefit of greater consistency across tasks than human ratings can achieve (Cushing Weigle, 2013). AES offers an efficient way to collect two independent scores per essay: one human score and one score derived by AES. In addition to objectivity, AES scores have the additional benefits of low cost, fast score reporting and reducing teachers' workloads. Starting with Project Essay Grade in the 1960s (Page, 1966), these advantages have led to the development of an array of applications based on automated essay scoring algorithms, including e-rater (Burstein et al., 1998), Intellimetric (Elliot, 2003) and Intelligent Essay Assessor (Foltz, Landaur, & Laham, 1999).

### 4.1.2 The validity of automated essay scoring

In test validation, validity is defined as "the degree to which evidence and theory support the interpretation of test scores entailed by their proposed use" (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). Since the emergence of AES, its validity has been much debated. Page and Petersen (1995) listed three general objections to automated scoring. The first objection is that computers will never be able to understand and/or appreciate language in the same way as humans do. Second, AES systems may not be able to detect "bad faith essays," that is, essays that are written with the intention to trick the rating system by anticipating the rating criteria, which results in unjustly high (or low) scores. The third objection implies that the linguistic (surface) features measured by automated systems do not represent what is intrinsically important in essay quality.

Using human judgment as a criterion

Attali and Burstein (2006) summarized the large body of previous studies that have attempted to validate the use of AES. The most common type of validation consists of the comparison of machine-human agreement to human-human agreement. Typically, these studies reported a high agreement between human qualitative judgments and automated measures of writing quality (e.g., Atalli & Powers, 2008; Crossley, Weston, Sullivan, & McNamara, 2011; Deane & Quinlan, 2008; Elliot, 2003). This indicates that the traits measured by AES are in fact characteristics of the writing construct (Deane & Quinlan, 2010).

However, a high agreement between AES scores and human scores alone does not provide adequate support for the validity of AES as an assessment of writing ability. The sole fact that AES measures certain characteristics of writing does not mean that AES covers the construct of writing. Essay length, for example, is known as an excellent predictor of text quality, and raters generally place higher value on longer essays. However, essay length in itself is not an indicator of text quality; instead it coincides with text quality. This measure will thus yield high agreement between human and computer scores without actually being a valid measure of text quality and therefore of writing ability (Attali & Burstein, 2006).

In addition, Clauser, Kane, and Swanson (2002) argued that expert judgment, as a single criterion, is problematic in assessing the validity of machine-scored responses because of the varying quality of human judgment. Even trained experts are known to differ in their judgments, and criteria are applied differently across raters and within raters across time, which results in flawed inter-rater and intra-rater reliability estimates. Furthermore, Williamson et al. (2012) stated that the scores obtained by humans and machines are expected to differ because of typical human rater problems (e.g., lack of consistency and fatigue) that do not occur in machine scoring. Attali & Burstein (2006) noted that inter-rater reliability prevents agreement with any other measure from rising beyond human-human agreement. This demonstrates the need to look beyond human judgment as a criterion when validating automated test score use (Chapelle & Chung, 2010). Over the last decade, this stance has been advocated in studies concerning the validity of automated scoring (e.g., Attali & Burstein; Bennet & Bejar, 1998; Ben-Simon & Bennett, 2007).

Additional approaches to the validity of AES

Williamson, Xi, and Breyer (2012) provided a framework of guidelines for the use of automated scoring. First, the authors presented an overview of recent developments in studies on automated scoring, and they noted that studies on AES have evolved from asking *whether* automated scoring *can* be done, to examining *how* it *should* be done. Subsequently, the authors stated that the process of collecting best practices for automated scoring is an ongoing process, and they reviewed a decade of conceptual validity frameworks for automated scoring.

The review of these validity studies revealed both a trend towards applying an argument-based approach to test validation and a trend to go beyond human-automated

score agreement (Williamson et al., 2012). Thoughts on the concept of validity have changed over the last decades. Evolving from a check on whether a test measures what it purports to measure, validity is regarded nowadays as the degree to which both theory and empirical evidence support the inferences and actions, based on test scores (Kane, 2006). This approach implies that the act of validation should be defined as an on-going process of gathering evidence that supports the intended interpretations of test scores. In this concept of validity, validation is only possible in a particular context (Kane, 2006). Moreover, the specific type of evidence to be collected in a validation study depends on the intended use of the test and the interpretation of the test results (Clauser et al., 2002).

In this trend towards argument-based validation, ideas about evidence that is best suited for the validation of AES scores are subject to change. In addition to the common approach of using agreement with human scores as a criterion, Yang, Buckendahl, Juszkiewicz, and Bhola (2002) discussed two alternative approaches to the validation of AES. The first approach focuses on the relation between automated scores and other measures of writing ability (or a similar construct). The second approach concerns the interpretability of AES output and the importance of understanding the underlying scoring processes of AES, instead of merely using them to achieve agreement with human scores. Although these last two approaches are arguably of greater importance to the validity of AES than are comparisons to human scores, they have remained unaddressed in studies on AES, as Attali & Burstein (2006) pointed out.

Currently, however, studies using validation approaches that go beyond agreement between human and automated scores are more common. The results of these studies have indicated that AES applications tap on the same construct of writing ability as evaluated by human raters. Moreover, AES applications provide judgments that can be validly interpreted as indicators of writing ability (Shermis, Burstein, Higgins, & Zechner, 2010, Attalli & Burstein, 2006).

### 4.1.3 Automated essay evaluation within primary education

The present study aims to explore the use of automated scoring techniques within the Dutch national assessment of writing ability in primary education (PPON: *Periodieke Peiling Onderwijsniveau,* Dutch National Assessment in Education). Within this large-scale assessment, the writing ability of Dutch pupils is evaluated halfway through their primary school career (i.e., grade 3, ages 8 and 9) and at the end of primary school (i.e., grade 6, ages 11 and 12). A set of writing tasks is used to collect a large sample of texts differing in format and genre. A pool of trained raters scores these essays according to several criteria of text quality. Based on these ratings, conclusions are drawn on the level of writing ability of both populations (i.e., grade 3 and grade 6) (Kuhlemeier, Van Til, Hemker, De Klijn, & Feenstra, 2013). An analysis of textual quality is performed in addition to the evaluation of writing ability (Van Til, Kuhlemeier, Hemker, & Keune, 2014).

Analyses of linguistic surface features

Critics of AES often put forth the argument that automated scoring techniques are only capable of scoring the surface features of texts, thus ignoring more meaningful, deeper text features (cf. Deane & Quinlan, 2010; Shermis, Burstein, Apel Bursky, 2013). However, in the evaluation of the written products of novice writers, surface features are meaningful indicators of writing ability, since the development of writing skills in the second half of primary education is generally considered concentrated at the surface level of language. The compositions of novice writers at this stage will therefore still be constrained by word use, sentence structure and and discourse structure (Abbott & Berninger, 1993, Crossley et al., 2011).

This theory is supported by the finding that it takes pupils a relatively long time to achieve a level of written proficiency that is similar to the oral proficiency they demonstrate at an earlier stage in their language development (Loban, 1976). Moreover, compared to spoken language, the written language produced by young children is shorter and of lower quality (Berninger & Swanson, 1994). It is commonly assumed that these differences are caused by the demanding cognitive process of translating thoughts into words (Silva, Sánchez Abchi, & Borzone, 2010). In addition, novice writers are likely to be hindered by the specific demands of producing written language, namely constructing a text to be read by a distant reader (Bereiter & Scardamalia, 1987). To ensure that this distant reader will be able to understand the writing, the writer has to adhere to numerous linguistic conventions (e.g., spelling rules and grammar), and he or she has to provide a coherent text with appropriate wording—usually without receiving feedback. Speaking, on the other hand, is a dynamic process in which the speaker interacts with an audience to ensure that the oral message is well received.

Regarding their cognitive model of writing development, Bereiter and Scaradamalia (1987) stated that the production of written language by writers in an early developmental stage, which is referred to as "knowledge telling," is hindered by the process of coding thoughts into written language. According to cognitive models of the writing process (Flower and Hayes, 1981; Hayes, 2012), which consist of the components *planning*, *translating, transcribing* and *reviewing*, developments in early writing skills will thus take place at the stage of *translating* ideas into words. In this stage, writers translate their thoughts and ideas into words, a process during which they need to take into account all the specific demands of written language in order to produce a text that is comprehensible to a targeted reader. It is assumed that this process places heavy demands on working memories of novice writers, thus constraining their written output (Flower & Hayes, 1981; Silva et al., 2010).

Given these characteristics of novice writing, the fact that AES is mainly suited to evaluating surface features underlines its ability to measure the textual features that are likely to differ in the written products of pupils at varying levels of ability, thus *enhancing* the validity of writing assessment instead of *threatening* it. In line with this stance, Attali and Powers (2008) stated that textual features, such as fluency, conventions and word choice,

are likely to be suitable measures for the assessment of novice writing on the one hand and on the other hand, apt measures for AES technology.

Because at the translation stage the writing ability of novice writers is still developing, their writing products are expected to be linguistically flawed at the surface level. Indeed, a recent study on the textual quality of writing products in primary education reported a relatively large number of flaws in spelling, grammar and punctuation (Van Til et al., 2014). These defects can influence the validity of automated scoring outcomes because misspelled words may not be recognized, ungrammatical sentence structures can prove impossible to parse, and lack of punctuation may lead to long sentences that will be incorrectly rated as complex structures. In the present study, the extent to which these flaws are present and the degree to which their presence influences automated evaluation will be examined before presenting further analyses.

Multi-trait evaluation of writing

Most AES tools are built by constructing an algorithm that generates an overall score based on separate features that are measured by the tool (e.g., word frequency and sentence length). These scores can be constructed such that they show high agreement with human essay scores. Hence, they may be deployed as a second, objective rating in addition to a human score. However, these overall scores do not give specific information about writing ability and are therefore less useful in a (diagnostic) assessment. Recently, however, automated essay *evaluation* (AEE) has evolved, which is an application of AES techniques designed to offer feedback on different aspects of writing ability, instead of holistic scores that approximate human scoring (cf. Shermis et al., 2013).

Arguably, the evaluation of texts by novice writers could benefit from an AEE approach, in which several aspects of writing ability would be evaluated separately. Because language learners are still developing their skills, individual novice writers may differ in their performance across different aspects of writing (Lee, Gentle, & Kantor, 2009). Holistic scoring is incapable of capturing specific weaknesses and strengths (Cushing Weigle, 2002); thus, analytical scoring—also referred to as multi-trait scoring—appears better suited to evaluating the writing ability of novice writers.

### 4.1.4 Automated analyses of text complexity in Dutch: T-Scan

Although an array of tools is available for the automated evaluation of texts written in English, considerably less work has been done on the development of scoring engines for writing in other languages. Because languages differ—to a greater or lesser extent—with respect to the characteristics that typically affect the evaluation of text quality (e.g., the structure of sentences and words), the AEE measures selected to evaluate writing in English cannot simply be applied to other languages. The specific features of English language that underlie these measures might contribute differently (or not at all) to the quality of writing in other languages, or they might not be present in the target language. Consequently,

building AEE-tools based on languages other than English requires a renewed process of identifying specific features that indicate writing ability in the target language.

Currently, no applications that are specifically designed to offer an extensive automated evaluation of writing in Dutch are available. In this study, therefore, the Dutch program T-Scan (Kraf, Van der Sloot, Pander Maat, Van den Bosch, & Kleijn, 2013) is used to explore the linguistic features of written products. T-Scan is designed to perform an elaborate analysis of text complexity, which in turn can serve to predict readability. T-Scan combines two NLP-applications for the analysis of Dutch language: Frog (Van den Bosch, Busser, Daelemans, & Canisius, 2011) is used for tokenizing, part-of-speech tagging and morphological parsing; Alpino (Van Noord, 2007) executes syntactic parsing within T-Scan. Also incorporated are word frequency data and annotated corpora of the written language. Using this set of applications, T-Scan (version March 2013) identifies 147 measures of text complexity, which are classified into eight different categories (Table 1).

Table 1. *Measures of Text Complexity in T-Scan*

| Category | Number of features | Text features measured (examples) |
|---|---|---|
| Word complexity | 24 | word length, word frequency |
| Sentence complexity | 30 | sentence length, sentence complexity |
| Information density | 12 | i.a. type/token ratio, content words, adverbials |
| Coherence | 17 | connectives, argument overlap, indefinite noun phrases |
| Concreteness | 12 | concrete nouns and adjectives |
| Personality | 23 | i.a. personal pronouns, action verbs, interrogatives |
| Parts of speech | 10 | i.a. prepositions, adverbs, interjections |
| Miscellaneous | 19 | i.a. present tense, copula, infinitives |

Because T-Scan was developed as a tool for the analysis of text complexity features as indicators of readability, its measures do not necessarily fit the analysis of writing skills. Although the linguistic complexity of a text is likely to indicate the proficiency of its writer, the development of writing skills may not necessarily be reflected in the measures of text complexity that are incorporated into T-Scan. For example, a proficient writer will be able to produce highly complex sentences, but is likely to try to adjust the level of complexity for the sake of the intended reader or the intended communicative goal. It might also be the case that the develop-mental pattern in some measures does not run parallel to an increasing score of complexity. As Deane and Quinlan (2010) pointed out, however, studies on readability and AEE usually include the same measures (i.e., fluency, complexity and accuracy), thus indicating a strong relation between the variables that reflect variation in both text complexity and writing quality.

Although AEE algorithms are not capable of mimicking the mental processes human raters use when judging writing competence, research has shown that AEE has strong correlations with human rating behaviour and that it is related to external measures of writing ability (Shermis et al., 2013). Because no valid definition of what constitutes good

writing is readily available, the development of AEE tools focuses on identifying linguistic features that represent the components of writing that are considered to reflect high proficiency. Hence, the present study aims to provide insight into the discriminant components of the writing ability of Dutch novice writers.

In this exploratory study, no preliminary assumptions on the relation between text complexity measures and writing ability are given. Instead, all text complexity measures are initially considered eligible as indicators of writing ability. Eventually, to be selected as an indicator of writing development, a measure should be related to other criteria of writing ability on the one hand, and should be interpretable as a measure of text quality on the other hand. To select the appropriate indicators, the relation between text complexity and writing ability is first explored by determining agreement between the text complexity measures used in T-Scan and two indicators of writing ability, namely grade level and human essay scores (Section 4.2). Based on the results of this quantitative study, a selection of measures is made and hypotheses on the relation between text complexity and writing ability are formulated. These hypotheses are then tested in a qualitative analysis of the relation between text complexity and writing ability (Section 4.3). Together, results of these analyses lead to a conclusion on the applicability of AEE within a large-scale assessment in primary education (Section 4.4).

## 4.2 Automated essay evaluation: exploring its applicability for novice writers

### 4.2.1 Introduction

The goal of this study is to explore the feasibility of automated essay evaluation (AEE) within a national assessment of writing ability in primary education. The study aims to identify valid descriptors of writing ability according the measures of text complexity offered by T-Scan. To do so, the inference that these measures of text complexity represent aspects of writing ability must first be supported. In other words, it has to be demonstrated that the scoring model (i.e., measures in T-Scan) corresponds to the construct (i.e., writing ability operationalized as text complexity) that is anticipated to explain the performance in question (Chapelle & Chung, 2010).

To demonstrate the correspondence between the scoring model and the construct, agreement between the values of the AEE measures derived from T-Scan and grade level as an external measure is analysed, as well as the agreement with human essay scores (cf. Yang et al., 2002). As Attali and Powers (2008) pointed out, performance on the linguistic features that underlie AEE measures increases progressively across grades. Agreement between AEE measures and grade level thus indicates the validity of these measures as predictors of writing ability, and agreement with expert human judgments of writing quality further supports their validity. In other words, analyses of agreement with grade levels and agreement with human judgments yields insight into the suitability of the measure to reflect known developmental trends and levels of text quality.

### 4.2.2 Research questions

The present study aims to answer the following research questions:

1. *What specific difficulties arise in analysing essays of novice writers? To what extent can these be overcome?*
2. *Which text complexity measures are suited to describe and evaluate the development in writing ability from mid-primary education (grades 3/4) to end primary education (grades 5/6)?*
    - 2.1. *Which measures are positively related with grade level? Which of these values differs significantly between grade 3 and grade 6?*
    - 2.2. *Which measures are positively related with (human) essay scores? Which of these values differ significantly between score level 1 and score level 4?*

This exploratory study aims at providing hypotheses on the use of linguistic measures of text complexity when evaluating writing, which will be tested by using a qualitative approach (Section 4.3). First, the expected difficulties encountered in analysing short texts that contain spelling errors and lack punctuation will be mapped and suggestions for future use will be offered. Furthermore, the findings of the analyses performed to determine agreement

between grade level and essay scores will serve as selection criteria for a set of measures related to writing ability. That is, the measures that correlate with these indicators of writing proficiency will be considered valid measures of writing ability. It is expected that a set of promising measures will be identified based on the analysis of the relation to grade level and human essay scores. Based on the outcomes, a prognosis will be made on the feasibility of using AEE to gain insight into both the writing ability of novice writers and the development thereof.

### 4.2.3 Method

**Data collection and materials**

Essays

Five Dutch primary schools representing different regions, school sizes, and denominations volunteered to participate in the study, which included 584 pupils, aged 8 to 12. Three different essay tasks were selected from the pool of tasks in the Dutch national assessment (Krom et al., 2004) covering a broad scope of communicative goals and text genres (Table 2).

Table 2. *Essay Tasks*

| Task | Description | Communicative goal | Text genre |
|---|---|---|---|
| A Tigors & Giraks | Finishing an adventurous story on two tribes | Narrative | Story |
| B Pookie | Writing a note describing a lost cat and requesting help | Descriptive/ Directive | Leaflet |
| C Yummie | Writing a letter to convince a company to accept an incomplete stamp card | Argumentative/ Persuasive | Letter |

Each pupil wrote either two or three essays by hand, and a total of 1,476 essays were collected. Of the three administered essay tasks, prompt C was selected to be evaluated in the present study. Consequently, 418 essays were included in the analysis (Table 3). Prompt C was selected because it requested a more or less formal letter directed to an unknown adult, which was intended to elicit full sentences instead of enumerations (cf. prompt B) and contained few stylistic elements or narrative elements (cf. prompt A), such as direct and indirect speech. Furthermore, since its communicative goal was argumentative, letters based on prompt C should feature intelligible, complete and correct reasoning in order to reach the writing goal—criteria that are likely be suitable for the evaluation of text quality based on text complexity measures. In contrast, quality criteria for narrative texts based on prompt A would be grippingness, originality and vivacity, that is, textual features that are difficult to define and hence to evaluate automatically.

Table 3. *Sample of Essays*

| | Grade 3 age 8/9 | Grade 4 age 9/10 | Grade 5 age 10/11 | Grade 6 age 11/12 | Total |
|---|---|---|---|---|---|
| Girls | 53 | 51 | 54 | 40 | 198 |
| Boys | 47 | 61 | 58 | 52 | 220 |
| **N essays** | **100** | **113** | **113** | **92** | **418** |

Human essay scores

To avoid the influence of hand writing on the essay scores and to facilitate logistics, all essays were retyped, maintaining layout, typos, and punctuation. Three aspects of writing ability were identified: content, structure and correctness. The procedure described by van den Bergh and Rijlaarsdam (1986) was adopted to compose a rating scale with anchor essays for each aspect per task. The result was nine rating scales in total, each with three anchor essays representing specific ability levels (Figure 1). Essay scores were composed of a list of analytical questions per aspect and an overall score per aspect, based on comparison with the anchor essays. Each essay was rated by a minimum of two trained raters randomly assigned from a pool of 13 raters. For the prompt selected in this study, inter-rater agreements of .87 (Content), .84 (Structure) and .77 (Correctness) were found (cf. Chapter 2).

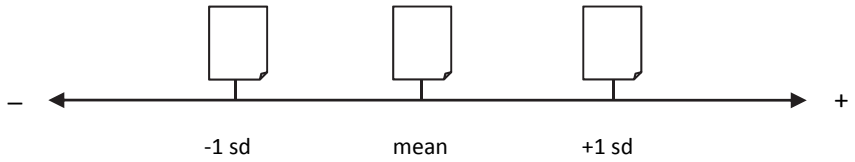



*Figure 1.* A rating scale with three exemplars.

Pilot with T-Scan

A small pilot analysis was conducted in which a random sample of 50 essays across grades was evaluated using T-Scan. This analysis revealed a negative influence of spelling errors and lack of punctuation on the validity of the complexity scores. Figure 2 shows two examples of essays that are flawed. Essay (A) is a typical, low-quality essay: a very short text containing a relatively large number of spelling errors and no punctuation. Essay (B) is severely flawed with respect to punctuation, but nonetheless demonstrates higher writing ability than (A), according to the human raters. The validity of the T-Scan output is influenced in different ways by flawed input.

First, words that are spelled wrongly *sometimes* (but not always) receive an incorrect part-of-speech tag. In (A) for example, the first instance of *spaaren* (correct form *sparen*: to collect) is incorrectly tagged as a noun, while the second instance is correctly tagged as an (unknown) verb. Incorrect part-of-speech tags influence all T-Scan measures

that involve the specification of word classes (e.g., the use of referential pronouns), whereas unknown words influence measures that are based on specific word properties (e.g., noun concreteness and word frequency).

Second, the lack of punctuation leads to seemingly long and complex sentences, causing "noise" in the complexity measures. In T-Scan, sentences are identified based on punctuation elements that mark the end of a sentence (i.e., period, exclamation mark and question mark). In the absence of such a mark, all words are processed as if they form one long sentence. Consequently, in addition to the apparent effect on sentence length, flawed or absent punctuation influences all measures in which the sentence is used as a unit of analyses. Of T-Scan's 147 measures, 28 were found to be influenced by sentence length. These measures are indicated in Appendix A.

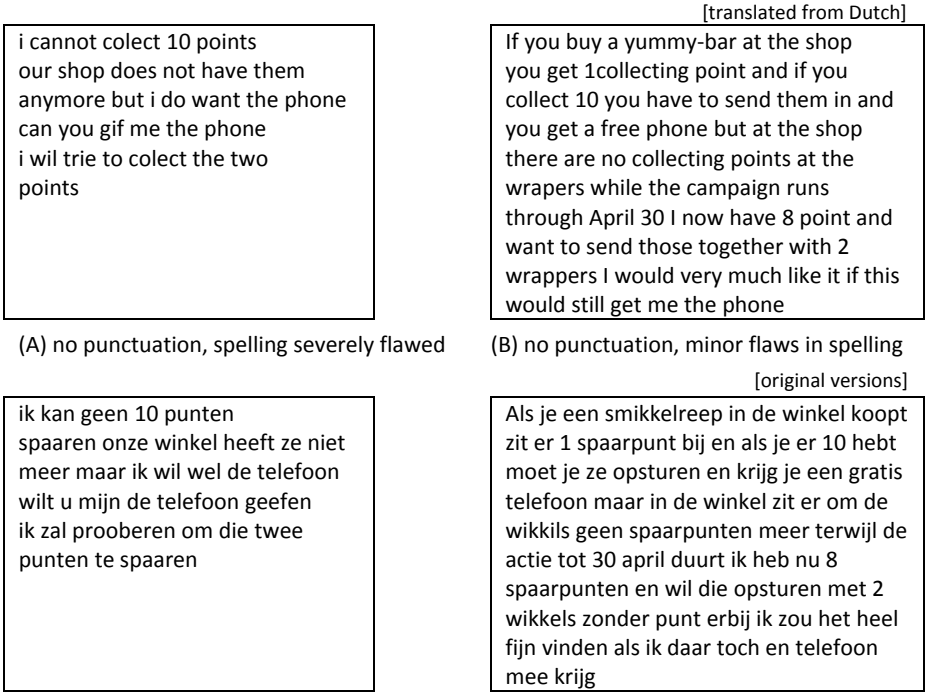



*Figure 2*. Examples of flawed essays.

The validity of almost every measure in T-Scan will, more or less, be at risk when flawed input is analysed. Because T-Scan is designed to conduct a complexity analysis of correct (adult) language, no information is currently available on the specific effects of flawed-input texts regarding the validity per complexity measure. Therefore, for the purposes of this study, flaws in spelling and punctuation were eliminated prior to the analysis using T-Scan. Hence, a valid analyses of text complexity was enabled despite the potential effects of flaws.

Essay editing

All essays were checked by one linguist, and if necessary manually edited in three ways. First, because the essay prompt elicited a letter format, most essays contained letter heads

and introductory and/or closing salutations. These components were removed because they normally do not consist of full sentences and are closed by a comma instead of a sentence stop (i.e., a period, exclamation mark or question mark) and would therefore influence measures of sentence complexity. Secondly, a total of 470 spelling errors was detected and corrected (Table 4). Finally, a total of 110 essays lacked punctuation almost entirely, which resulted in texts that were impossible to parse grammatically. In these essays, the minimum number of sentence stops (i.e., period or question mark) required to produce an intelligible text was added. After this editing, all essays were successfully processed by T-Scan in one batch.

Table 4. *Corrected Spelling Errors*

| Error category | Example | Correct form | N |
|---|---|---|---|
| Verb | ik hed; ik deedt | ik heb; ik deed | 112 |
| Duplication of vowels | telefon; darom; heele | telefoon; daarom; hele | 58 |
| Duplication of consonants | tellefoon | telefoon | 46 |
| Similar sounding consonants | egt; overtuichd; goet | echt; overtuigd | 54 |
| Spacing | spaar punten; er bij | spaarputen; erbij | 89 |
| Other | ackoort; niewe; mischien | akkoord; nieuwe; misschien | 111 |

**Analyses**

The collected output from the analysis using T-Scan consisted of the values per measure of text complexity per essay. This output was used to analyse both the relationship to grade level of the pupil and the human scores on text quality.

Relation with grade level

In 421 essays, the mean values per T-Scan measure were computed for each grade (3 to 6). Based on these values, 35 measures for which the values appeared to increase or decrease systematically according to grade level were selected. Next, as an evaluation of their usability as indicators of writing ability, the selected measures were evaluated for statistical quality.

First, the distribution of values was evaluated. Measures for which too few observations were found across groups were excluded from further analysis (marked A in Table 7). For measures that were assigned a value of 0 in more than 10 per cent of cases and for measures that contained extreme values, additional analysis were performed to determine whether this influenced the score distribution across grades. If this was indeed the case, all values of 0 were removed from the analysis (marked B in Table 7).

Next, the Shapiro–Wilk test was used to determine whether all measures showed normal distribution per grade level. For measures for which no normal distribution was found, the non-parametric Wilcoxon rank sum test was applied to test the significance of the difference between values for grade 3 and grade 6 (marked C in Table 7). For all other measures, the significance was tested using an unpaired two-sample t-test.

Relation with human essay scores

For 236 essays, the human scores on writing ability, which were assigned by two trained raters, were available for content, structure and correctness. These scores were based on the un-edited versions of the essays. Using the overall score per essay (i.e., the sum of the scores per aspect), percentile scores were computed (Table 5), and the essays were divided into four groups of proficiency across grade levels (Table 6).

The statistical quality of these measures was evaluated. First, the distribution of values was evaluated, and the T-Scan measures for which too few observations were found were excluded (marked A in Table 8). The influence on the score distribution across grades was checked in measures that were assigned a value of 0 or 1 in more than 10 per cent of cases (i.e., they were either absent or present in a large proportion of essays) and in measures that contained extreme values. If an influence was indeed found, all values of 0 or 1 were removed from the analysis (marked B and C, respectively, in Table 8). Second, the Shapiro-Wilk test was applied to all measures to test for normality. If no normal distribution was found, the non-parametric Wilcoxon rank sum test was used (pairwise) to test significance (marked D in Table 8). For all other measures, the significance of the difference in values for score level 1 and score level 4 was tested using an unpaired two sample t-test.

Table 5. *Human Essay Scores in Percentile*

| **Percentile** | **p10** | **p25** | **p40** | **p50** | **p60** | **p75** | **p90** |
|---|---|---|---|---|---|---|---|
| Mean sum score | 10.00 | 16.00 | 20.37 | 23.33 | 26.00 | 28.50 | 32.07 |

Table 6. *Essays per Score Level*

| **Score level** | **1** | **2** | **3** | **4** | |
|---|---|---|---|---|---|
| Percentiles | ≤p25 | >p25 - p50 | >p50 - p75 | >p75 | **Total** |
| N essays grade 3 | 32 | 16 | 13 | 5 | 66 |
| N essays grade 4 | 19 | 22 | 13 | 23 | 77 |
| N essays grade 5 | 8 | 18 | 25 | 20 | 71 |
| N essays grade 6 | 2 | 2 | 9 | 9 | 22 |
| **N essays total** | **61** | **58** | **60** | **57** | **236** |
| N girls | 20 | 30 | 30 | 33 | 113 |
| N boys | 41 | 28 | 30 | 23 | 122 |
| unknown | 0 | 0 | 0 | 1 | 1 |

### 4.2.4 Results

Relation with grade level

Table 7 presents the results of the grade level analysis of 35 T-Scan measures. The mean values per grade level are given for each measure, followed by the significance levels of the difference in mean values for grade 3 (age 8/9) and grade 6 (age 11/12). The codes A to C under Comment provide the specific aspects of each measure: exclusion based on the number of observations (A); removal of values 'zero' (B); use of a non-parametric test (C).

Where relevant, boxplots of these results are given in Appendix B to illustrate the mean values, standard deviations and outliers per grade level.

Table 7. *Relation between Grade Level and 35 Complexity Measures in T-Scan*

| | | Means per grade level | | | | | |
|---|---|---|---|---|---|---|---|
| | **Type**[1] | 3 | 4 | 5 | 6 | **Significance**[2] | **Comment**[3] |
| **Measure** | | (n=101) | (n=113) | (n=113) | (n=94) | 3 vs. 6 | |
| **Word complexity** | | | | | | | |
| word length (n letters) | m(w) | 3.899 | 3.994 | 3.996 | 4.057 | *** | |
| word complexity (n morphemes) | m(w) | 1.194 | 1.224 | 1.236 | 1.242 | *** | |
| word frequency (50%) | p | 0.563 | 0.557 | 0.545 | 0.535 | *** | |
| **Sentence complexity** | | | | | | | |
| subordinate clauses | d | 17.282 | 22.120 | 26.061 | 26.571 | *** | |
| relative clauses^ | d | 0.578 | 1.061 | 1.857 | 2.214 | NA | A |
| dependency length subject-verb | m(t) | 0.954 | 1.320 | 1.551 | 1.803 | *** | |
| dependency length | m(t) | 1.355 | 1.539 | 1.676 | 1.798 | *** | |
| content words | m(c) | 2.543 | 2.590 | 2.661 | 2.820 | *** | |
| clauses | d | 165.375 | 157.582 | 151.640 | 147.296 | *** | |
| sentence length | m(t) | 9.755 | 11.596 | 12.667 | 12.337 | *** | C |
| **Coherence** | | | | | | | |
| contrastive connectives | d | 39.791 | 34.519 | 33.006 | 26.073 | *** | |
| comparative connectives | d | 11.840 | 11.567 | 7.139 | 7.898 | NA | A |
| temporal connectives | d | 3.849 | 5.327 | 5.254 | 7.601 | NA | A |
| enumerative connectives | d | 45.107 | 41.442 | 35.894 | 35.030 | ** | |
| causal connectives | d | 47.392 | 42.120 | 44.720 | 38.080 | - | B |
| referential pronouns | d | 71.416 | 69.400 | 66.291 | 56.499 | ** | C |
| argument overlap (lemmabuffer) | d | 83.980 | 78.342 | 78.343 | 72.141 | * | C |
| **Information density** | | | | | | | |
| adverbials | m(c) | 0.871 | 0.937 | 1.022 | 0.983 | * | |
| type/token-ratio | r | 0.707 | 0.714 | 0.694 | 0.679 | * | |
| **Personality** | | | | | | | |
| noun concreteness (strict) | p | 0.771 | 0.727 | 0.727 | 0.664 | *** | C |
| noun concreteness (broad)^ | p | 0.798 | 0.748 | 0.762 | 0.691 | *** | C |
| personal references^ | d | 155.033 | 155.324 | 148.560 | 145.747 | - | C |
| personal pronouns (1st person)^ | d | 113.413 | 108.800 | 105.642 | 102.667 | ** | |
| personal pronouns (2nd person)^ | d | 131.023 | 129.301 | 125.606 | 126.973 | - | |
| personal pronouns (3rd person) | d | 143.118 | 138.420 | 131.220 | 131.847 | * | |
| **Miscellaneous** | | | | | | | |
| prepositions | d | 47.818 | 59.883 | 62.865 | 80.365 | *** | |
| imperatives^ | d | 8.342 | 6.671 | 5.703 | 5.105 | NA | A |
| questions^ | d | 9.330 | 9.059 | 5.737 | 4.775 | NA | A |
| present verbs | d | 148.729 | 134.544 | 125.941 | 117.696 | *** | |
| modals | d | 52.482 | 46.260 | 44.666 | 39.414 | *** | B |
| time verbs^ | d | 0.130 | 0.159 | 0.203 | 0.212 | NA | A |
| copula^ | d | 23.590 | 16.657 | 16.176 | 16.070 | NA | A |
| present particle^ | d | 0.000 | 0.136 | 0.167 | 0.500 | NA | A |
| infinitive^ | d | 31.647 | 37.548 | 37.946 | 43.386 | NA | A |
| predictability | m(s) | 1.281 | 1.412 | 1.591 | 1.565 | *** | C |

[1] computation method for measure: m=mean per word(w), clause (c), text (t); d=density (value per 1000 words); p=proportion

[2] significance level of difference in value for grades 3 and 6: *** p < 0.001, ** p <0.01, * p<0.05, - p>0.05 (NA: too little observations)

[3] A:not enough observations; B:all values '0'removed; C:no normal distribution, non-parametric test applied

^discarded from further analysis (cf. Table 8)

Relation with score level

Based on the outcomes of the grade level analysis, 11 T-Scan measures were discarded (denoted by ^ in Table 7). Table 8 presents the results of the essay score analysis for 24 T-Scan measures. For each measure, the mean values per ability group within grade level are given (cf. Table 6), followed by the significance levels of the mean values for score level 1 and score level 4. The codes A to D listed under Comment provide the specific aspects of each measure: exclusion based on the number of observations (A); removal of all values 0 (B); removal of all values 1 (C); and use of a non-parametric test (D). Where relevant, boxplots of these results are given in Appendix C, illustrating the mean values, standard deviations and outliers per score level.

Table 8. *Relation between Score Level and 24 Complexity Measures in T-Scan*

| | | Means per score level | | | | | |
|---|---|---|---|---|---|---|---|
| **Measure** | **Type[1]** | 1 (n=61) | 2 (n=58) | 3 (n=61) | 4 (n=59) | **Significance[2]** 1 vs. 4 | **Comment[3]** |
| **Word complexity** | | | | | | | |
| word length | m(w) | 3.848 | 4.004 | 4.019 | 4.067 | *** | |
| word complexity | m(w) | 1.190 | 1.231 | 1.237 | 1.241 | *** | |
| word frequency (50%) | p | 0.562 | 0.560 | 0.548 | 0.532 | * | |
| **Sentence complexity** | | | | | | | |
| subordinate clauses | d | 18.365 | 19.054 | 25.839 | 27.145 | ** | D |
| dependency length subject-verb | m(t) | 0.993 | 1.341 | 1.509 | 1.734 | *** | D |
| dependency length | m(t) | 1.394 | 1.543 | 1.690 | 1.799 | *** | D |
| content words | m(c) | 2.615 | 2.628 | 2.649 | 2.806 | * | D |
| clauses | d | 161.977 | 156.032 | 153.140 | 147.411 | ** | |
| sentence length | m(t) | 10.751 | 12.016 | 11.584 | 12.777 | ** | D |
| **Coherence** | | | | | | | |
| contrastive connectives | d | 46.172 | 46.231 | 38.599 | 31.522 | *** | B |
| comparative connectives | d | 17.278 | 6.893 | 7.875 | 7.589 | NA | A |
| temporal connectives | d | 4.705 | 4.414 | 4.992 | 6.036 | NA | A |
| enumerative connectives | d | 35.058 | 26.750 | 29.847 | 32.658 | - | D |
| causal connectives | d | 32.798 | 24.040 | 32.560 | 29.336 | - | D |
| referential pronouns | d | 33.494 | 32.434 | 32.982 | 28.326 | - | D |
| argument overlap (lemmabuffer) | d | 78.607 | 79.672 | 77.509 | 71.421 | - | D |
| **Information density** | | | | | | | |
| adverbials | m(c) | 0.899 | 0.974 | 0.958 | 1.055 | * | |
| type/token-ratio | r | 0.733 | 0.704 | 0.698 | 0.686 | ** | |
| **Personality** | | | | | | | |
| noun concreteness (strict) | p | 0.728 | 0.758 | 0.744 | 0.740 | * | C |
| personal pronouns (3rd person) | d | 139.532 | 138.609 | 131.068 | 125.272 | * | D |
| **Miscellaneous** | | | | | | | |
| prepositions | d | 48.866 | 57.073 | 67.146 | 68.395 | ** | |
| present verbs | d | 142.955 | 129.366 | 127.353 | 122.496 | ** | |
| modals | d | 41.814 | 45.091 | 39.354 | 38.627 | - | D |
| predictability | m(s) | 1.346 | 1.441 | 1.505 | 1.615 | *** | D |

[1] computation method for measure: m=mean per word(w), clause (c), text (t); d=density (value per 1000 words); p=proportion, r=ratio

[2] significance level of difference in values between score levels 1 and 4: *** p < 0.001, ** p <0.01, * p<0.05, - p>0.05

[3] A: not enough observations; B: all values '0'removed; C: all values '1' removed; D:no normal distribution, non-parametric test applied

Discarded measures

For a large proportion of the 35 selected measures that showed either an increase or a decrease with grade level and score level, a significant difference was found between values for grade 3 and 6 and/or score level 1 and 4 (cf. Table 7 and Table 8). However, 16 of these measures were discarded as potential measures of writing ability, based on one of three criteria. First, a number of measures were absent in most of the texts written by novice writers (e.g., relative clauses). Second, several measures evaluated content aspects of the texts, which are likely to be sensitive to specific aspects of the writing task, such as genre, audience and topic (e.g., noun concreteness). Finally, two measures proved superfluous because other measures of similar quality were used to evaluate the same feature. Table 9 provides an overview of the discarded measures.

Table 9. *Overview of Discarded Measures*

| **Discarded measure** | **Reason** |
|---|---|
| relative clauses<br>imperatives<br>questions<br>time verbs<br>copula<br>present particles<br>infinitives | irrelevant: mostly absent in novices' written products |
| predictability<br>present verbs<br>modals<br>personal references<br>personal pronouns<br>noun concreteness<br>content words | invalid: content measures, sensitive to characteristics of essay prompt |
| word frequency (log)<br>dependency length (subj-verb) | superfluous: other measures* of same feature available |

* Word frequency, dependency length (overall)

Selected measures of writing ability

Based on the agreement between T-Scan's measures of complexity and both grade level and human essay score, as reported in Table 7 and Table 8, a total of 13 complexity measures were identified that appear to be valid indicators of writing ability. These measures cover three textual levels: word, sentence and paragraph/text. Using these measures, word use and—to a lesser extent—sentence complexity and coherence of written text can be evaluated. Table 10 presents the selected measures, and specifies their relation with writing ability by stating whether the values increased or decreased with indicators of increasing writing ability.

Table 10. *Selected Complexity Measures for an Automated Evaluation of Writing Ability*

| | Measure | Description | Relation to writing ability[1] |
|---|---|---|---|
| WORD | | | |
| **Lexical complexity** | word length | number of letters per word (average) | + increases with ability |
| | word complexity | number of morphemes per word (average) | + increases with ability |
| | word frequency (50%) | proportion of words that overlap with 50% most frequent words of word frequency list[2] | - decreases with ability |
| **Lexical richness** | adverbials | number of adverbials per clause (average) | + increases with ability |
| | type/token-ratio | number of different instances (tokens) per total number of lemmas (types) (average) | - decreases with ability |
| | prepositions | number of prepositions (per 1000 words) | + increases with ability |
| SENTENCE | | | |
| **Sentence complexity** | subordinate clauses | number of subordinate clauses (per 1000 words) | + increases with ability |
| | dependency length | distance between sentence components[3] (average) | + increases with ability |
| | clauses | number of clauses (per 1000 words) | - decreases with ability |
| | sentence length | number of words per sentence (average) | + increases with ability |
| TEXT | | | |
| **Coherence** | argument overlap | overlap between (lemmatized) arguments[4] in preceding 10 words (per 1000 words) | - decreases with ability |
| | referential pronouns | number of 3rd person personal/possessive pronouns and demonstrative pronouns (per 1000 words) | - decreases with ability |
| | connectives[4] | number connectives (of different categories[5]) (per 1000 words) | - decreases with ability |

[1] Increase/decrease with increasing writing ability

[2] Staphorsius, 1994

[3] subject-verb; direct object-verb; object-verb; verb-preposition; determiner-noun; preposition-noun; finite verb-main verb; subordinate conjunction-finite verb subordinate clause; coordinating conjunction-conjunct head; subordinate conjunction-main verb; noun-subordinate clause

[4] arguments: pronouns (excl. demonstratives); proper nouns; nouns; main verbs

[5] comprised measure: T-Scan provides separate measures per category

[6] temporal; enumerative; contrastive; comparative; causal

### 4.2.5 Discussion

In the foregoing section, the applicability of automated essay evaluation (AEE) to the texts of novice writers was explored. For this purpose, T-Scan, an application used to analyse the linguistic complexity of written Dutch was employed. To evaluate the extent to which text complexity measures provided by T-Scan indicate writing ability, the outcomes of this program were related to two other indicators of writing ability. First, the agreement between T-Scan measures and grade level was evaluated, providing insight into the development of complexity across grades. In addition, agreement with human essay scores was determined, indicating the relation between text complexity and text quality. Hence, the study aimed to identify measures that are meaningful descriptors of writing ability.

Agreement with grade level and human essay score

Of the 147 measures of text complexity in T-Scan, 35 measures that appeared to behave as a function of grade level (i.e., they either systematically increased or decreased with increasing grade level) were selected. For these measures, the significance of differences in values for

grade 3 and grade 6 was tested in 421 essays. Table 7 provides the results of this analysis. In 21 measures, a significant difference was found between essays written in grade 3 and essays written in grade 6, which indicates that these measures serve well as a predictor of the developmental aspects of writing ability. For the remaining 14 measures, either too few observations were present in essays written by the novice writers (n=9), or no strong relation with grade level was found (n=5).

Agreement with human essay scores was evaluated in a selection of 24 measures. Based on the essay scores, a set of 239 essays was divided into score levels 1 to 4. Subsequently, tests were conducted to determine whether the differences in values between score level 1 and score level 4 were significant. Table 8 summarizes the results of this analysis. In 17 out of 24 measures, a significant difference was found in the values of essays of low quality and essays of high quality, indicating that these measures serve well as a predictor of text quality. In the remaining measures, either no strong relation with essay score was found (n=7) or too few observations were present (n=1).

Although the measures presented in Table 10 differentiate between groups of pupils of different ages and abilities, they require further investigation of their outcomes to determine their validity as indicators of writing ability. This is especially true for measures that are composed of different instances (words) of the specific category, namely prepositions, referential pronouns and different categories of connectives. These individual instances might show a different relation to writing ability. For example, the measure "referential pronouns" is based on a fixed collection of words that possibly show a different developmental pattern. Similarly, separate measures for connectives require further investigation. T-Scan incorporates extensive lists of connective elements that are clustered by function (e.g. contrastive or enumerative). However, within these categories, only a small number of connectives is used by novice writers. Furthermore, even within the categories, the use of some connectives appears indicative of low ability (e.g., contrastive *maar*: but), whereas the use of other connectives is likely to indicate a high proficiency level (e.g., contrastive *desondanks*: nevertheless). This finding is reflected in the results presented in Table 7 and Table 8, which show that the values of some connectives demonstrated a capricious pattern. Further investigation of the use of connectives per grade level is expected to provide further insight into the interpretability of connectives as a measure of writing ability (Section 4.3).

Construct representation of selected measures

Based on the analyses of both grade level and score level, Table 10 presents a selection of linguistic feature variables that are indicative of writing ability and that can be linked to aspects of writing. In order to explore the extent to which these measures cover the construct of writing, Ben-Simon and Bennett (2007) listed a series of commonly cited characteristics of good writing, which serves to represent theoretical meaningfulness in writing assessment. These characteristics are grouped into several aspects of writing ability,

as shown in Table 11. The selected T-Scan measures presented on the right represent part of the characteristics grouped under rhetorical structure (i.e., text structure in support of the rhetorical goal), vocabulary, and syntax and grammar; content and style are not covered.

Table 11. *Conceptual Links between Characteristics of Writing* (cited in Ben-Simon & Bennett, 2007) *and Selected Measures from T-Scan*

| **Writing ability** | | **T-Scan** | |
|---|---|---|---|
| **Aspect** | **Characteristic** | **Aspect** | **Measure** |
| Content | Relevance<br>Richness of ideas<br>Originality<br>Quality of argumentation | - | - |
| Rhetorical structure | Paragraphing<br>Coherence<br>Cohesion | Coherence | referential pronouns<br>argument overlap<br>connectives |
| Vocabulary | Richness<br>Register<br>Accuracy<br>Appropriateness to written language | Lexical richness<br>Lexical complexity | adverbials<br>type/token-ratio<br>prepositions<br>word length<br>word complexity<br>word frequency |
| Syntax and grammar/ Mechanics | Sentence complexity<br>Syntactical accuracy<br>Grammatical accuracy<br>Spelling | Sentence complexity | subordinate clauses<br>dependency length<br>clauses<br>sentence length |
| Style | Clarity<br>Fluency | - | - |

The conceptual links illustrated in Table 11 correspond with the common finding that AEE is capable of evaluating mainly "low level skills" (i.e., skills on the micro- and meso-text level), whereas "high level skills" (i.e., skills on the macro-text level) are either under-represented or not represented (Deane & Quinlan, 2010). Indeed, Table 11 shows that no aspects of writing that concern the content or style of the written product (e.g., providing sound arguments) are covered by the selected measures in T-Scan, whereas aspects concerning (local) rhetorical structure, vocabulary and syntax are amply represented. Furthermore, the results illustrated that AEE is largely based on frequency counts, which are informative when analysing writing development, but as Bereiter (1980) observed, "the variables they look at seem unrelated to purposes of writing instruction."

The underrepresentation of the writing construct is an anticipated shortcoming of AEE, because low level skills are expressed by certain measurable linguistic surface features (e.g., number of subordinate clauses), whereas high-level skills are usually not reflected in independently measurable units. In other words, content or style features are mostly "umbrella concepts" that are difficult to reduce to quantifiable units. Performance of higher level skills, however, depends upon the mastery of lower level skills that are needed to achieve high fluency in text production (Deane, 2013). Hence, AEE provides a measure of basic writing skills, which in turn can be used to predict the ability to perform high-order

skills (Deane, 2013). This notion is supported by Lee et al. (2009), who showed that, on average, language components are more predictive of human scores than are features of development and organization.

Linguistic features and writing ability

The study presented in this section identified several features of text complexity that are indicative of writing ability. A selection of these features and their relation to writing ability are presented in Table 10. On the word level, agreement between writing ability and the use of long, complex and infrequent words was found, which led to the conclusion that a high level of lexical complexity indicates a high level of writing ability. Similarly, the use of adverbials and prepositions increases with writing ability, indicating a relation between lexical richness and proficiency in writing. These findings support the notion that texts written by skilled writers demonstrate complex and diverse word usage.

Type/token-ratio (TTR), however, was found to decrease, not increase, with writing ability. TTR is computed by dividing the number of different words by the total number of words produced, and it is considered a measure of diversity in spoken or written language. This diversity is believed to increase with age, thus illustrating the growth of a child's vocabulary. As Richards (1987) pointed out, however, text length (i.e., the total number of words produced) increases the chance of word repetition within the utterance, resulting in a negative, instead of a positive, correlation between language development and TTR. Given this relation between TTR and text length, the former can arguably be interpreted as a measure of lexical *fluency* instead of *diversity*. TTR together with the use of adverbials and prepositions reflects the ability to produce a fluent text that is rich in information about time, place, manner, and so on. Hence, instead of lexical diversity, these measures are interpreted as indicators of lexical richness.

On the sentence level, the proportion of subordinate clauses, the overall dependency length and sentence length increased with writing ability, whereas the density of clauses decreased. These findings indicate that skilled writers produce sentences that are longer, more complex, and consist of longer clauses. Finally, on the text level, the density of referential pronouns and connectives and the overlap between arguments all decrease with writing ability. Skilled writers, in other words, seem to use relatively fewer (explicit) cohesive features when producing a text.

The relation between text complexity and writing ability found in the present study on Dutch novice writing, largely align with the results of analyses of the automated evaluation of essays written in English. For example, the relation between lexical complexity and lexical diversity is supported by results from Crossley et al. (2011), who reported the use of increasingly sophisticated words as grade level increased. In a study by McNamara et al. (2010), lexical diversity and word frequency were found to be the most predictive features. Lee et al. (2009) reported a positive relation among vocabulary level, word length and human scores, whereas TTR was found to be negatively correlated to essay quality.

In addition to endorsing the results of word complexity measures, the findings of the aforementioned studies support the findings of sentence complexity measures. Both Crossley et al. (2011) and McNamara et al. (2010) reported an increase in syntactic complexity with grade level. The relation between text complexity and sentence complexity found in the present study is further supported by the literature on writing development. Hunt (1966) stated that the construction of long sentences indicates high writing ability and that long sentences are more complex because they comprise more clauses. Hence, the low density of clauses indicates high writing ability, which corresponds to the present finding that clause density decreases with writing ability. Furthermore, Silva et al. (2010) stated that "syntactic maturity" develops with age and that the length of clauses and phrases increases with grade level, as does the use of subordinate clauses. An increase in the number of subordinate clauses is commonly found to be indicative of writing ability, see for example Hunt (1966) and the results of the Dutch national assessment (Van Til et al., 2013). However, Loban (1976) reported that the increase in the use of subordinate clauses levels off after grade 8, indicating that the use of subordinate clauses seems indicative of early writing development only.

Regarding the use of connectives, the present finding that the overall use of cohesive features decreased with grade level is supported by Crossley et al. (2011). Thus, although a coherent text indicates high writing ability (Sanders & Schilperoord, 2006; Sanders et al., 1996), coherence seems achieved by the use of relatively fewer cohesive features as writing ability increases. However, further analyses should reveal whether this *decrease* in density is in fact an effect of *increase* in text length because an increase in text length is expected to be caused mainly by an increase in content words instead of function words. Furthermore, the literature on language development indicates that coherence relations differ in complexity, which is reflected in the order of their acquisition (Spooren & Sanders, 2008; Evers-Vermeul & Sanders, 2009). In spoken language, additive relations (e.g., and) emerge before causal relations (e.g., because), and positive relations (e.g., and) emerge before negative relations (e.g., but) (Evers-Vermeul & Sanders, 2009). This acquisition order is based on the conceptual and syntactical complexity of connectives and results in the acquisition of both negative additive and positive causal connectives before negative causal relations.

In the light of the above, a qualitative analysis is needed to determine whether all individual cohesive devices demonstrate a similar relation to writing ability. The present results indicated that the use of cohesive devises cannot be readily interpreted as a measure of writing ability. Instead, specific cohesive elements are expected to demonstrate individually different relations to writing ability, based on their conceptual and syntactical complexity.

Finally, comparison of the results in Table 7 and Table 8 (i.e. the relation between text complexity and both grade level and score level) lends strength to the assumption that within this population, an increasing level of text complexity indicates higher writing proficiency. In all measures, a similar relation to both indicators of writing ability was found,

that is, measures that increase with grade level show an increase in score level, and vice versa. However, this relation will not hold for all populations because skilled writers are *able* to produce complex sentences, but can *choose* to produce shorter sentences, depending on their intended goal and audience. Less proficient writers, on the other hand, are likely to lack insight into restricting the complexity of their texts. Indeed, in a study on pupils in grades 4 through 12, Deane and Quinlan (2010) found an *increase* in sentence complexity with grade level, but a *decrease* with score level, indicating that increasing complexity does not always relate to increasing ability. Hence, the present findings indicated that writers in grade 6 are still developing text complexity and have not yet reached the point where they choose to confine complexity for the sake of clarity and readability.

**Conclusion**

The findings of the quantitative study performed in this section indicate fairly strong associations between essay feature variables measured by T-Scan and other indicators of writing ability (i.e., grade level and human ratings). However, the fact that in various measures of text complexity no relation with writing ability was found indicates that features of text complexity are not necessarily aligned with features of text quality (cf. McNamara et al., 2010). With respect to the measures that did correlate with grade level and essay scores, further analysis of the writing products is needed to gain insight into their specific relation to writing ability. Hence, in the following section, the validity of linguistic features of text complexity as measures of writing ability is evaluated.

## 4.3 Text complexity and writing proficiency: A qualitative analysis

### 4.3.1 Introduction

The validity of a test depends not only on its design—based on the *intended* use—but also on the *actual* interpretation of its scores (cf. Kane, 2006). Hence, understanding the measures underlying AEE is key in a valid interpretation of the scores provided by this method. However, most research of AEE occurs in the commercial sector, where efficiency-oriented research is typically not aimed at a better understanding of the underlying constructs (Chapelle, 2003; Chapelle, 2010). Hence, the techniques and procedures that underlie AEE remain unknown to language assessment professionals, who consequently maintain a sceptical attitude towards the validity of this approach (Chapelle, 2010). To improve the understanding of writing development and writing quality, the techniques and procedures that underlie AEE methods need to be transparent (Lee et al., 2009; Cushing Weigle, 2013; Ramineni & Williamson, 2013).

In general, text features have been implicated to varying degrees in different populations. Therefore, corpus analyses can help map the dimensions of writing ability and developmental change (Deane & Quinlan, 2010). By analysing the AEE outcomes in a corpus of essays, group differences across particular traits can be characterised, and different ability levels within the population can be illustrated. Hence, AEE can provide strong, corpus-build anchors to clarify what we mean by writing quality (Deane & Quinlan, 2010).

The goal of this section is to explore further the applicability of AEE within a national assessment in primary education. For this purpose, the validity and interpretability of a set of text complexity measures as indicators of writing ability is evaluated by analysing the specific features underlying the measures used in T-Scan. In addition, writing scales are constructed per cluster of measures, which illustrate the characteristics of texts written by pupils with differing abilities at different developmental stages. Based on these analyses, a conclusion will be drawn regarding the usability of the selection of T-Scan measures, based on their validity and interpretability.

### 4.3.2 Research questions

In the previous section, a quantitative study was performed in which a selection of text complexity measures was made (cf. Section 4.2, Table 10). Based on their relation to both grade level and human essay scores, these measures were assumed indicative of writing proficiency. With respect to the development in writing ability between grades 3 and 6, the following relations between complexity measures and writing ability are assumed:

a. Lexical complexity and lexical richness increase with increasing writing ability.
b. Sentence complexity increases with increasing writing ability.
c. The use of conceptually and syntactically unelaborate cohesive elements decreases with increasing writing ability.

These relations infer that as writing ability develops, pupils produce words that are longer, more complex and less frequent, and that their word use is diverse and reflect high fluency. Regarding the sentence level, it is expected that more subordinate clauses will be used and that the length of clauses and sentences will increase. Lastly, fewer additive connectives and personal referential pronouns will be used, and relatively less overlap between arguments will be produced. In the present section, these assumptions are qualitatively evaluated by means of the following research questions:

1. *To what extent are the selected measures valid indicators of writing ability?*
   - *1.1 To what extent do different instances of composite measures indicate a similar relation to writing ability?*
   - *1.2 Based on the analysis of outliers per measure, which factors that influence the validity of the measures can be identified?*
2. *To what extent are exemplar essays chosen by means of the selected measures interpretable as a developmental scale of writing ability?*

### 4.3.3 Method

Essays

Of a total of 1,476 essays written in response to three different writing tasks (A, B and C), all essays based on prompt C were selected to evaluate AEE (cf. Section 4.2.3). Hence, 421 essays written by pupils from grades 3 through 6 were included in the analyses. Table 12 shows the numbers of essays per group. Prompt C elicited formal, argumentative letters with an average length of 60 words. Figure 3 shows the average number of words per essay for each grade level.

Table 12. *Sample of Essays*

| | **Grade 3 Age 8/9** | **Grade 4 Age 9/10** | **Grade 5 Age 10/11** | **Grade 6 Age 11/12** | **Total** |
|---|---|---|---|---|---|
| Girls | 53 | 51 | 54 | 40 | 198 |
| Boys | 47 | 62 | 59 | 52 | 220 |
| N essays | 100 | 113 | 113 | 92 | 418 |

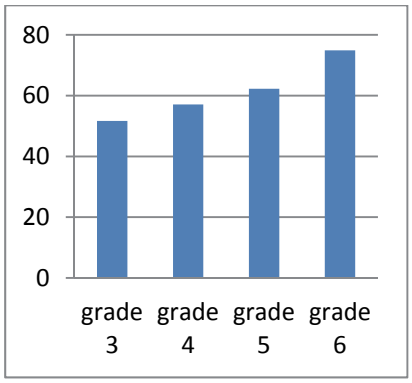



*Figure 3.* Average number of words per essay, per grade level.

Selection of measures

A selection of text complexity measures believed to be indicative of writing ability is used (Table 13). This selection is based on the agreement of T-Scan measures with two other indicators of writing ability, namely grade level and score level (cf. Section 4.2). These measures were evaluated in various ways, as described below.

Table 13. *Selected Measures for an Automated Evaluation of Writing Ability*

| | **Measure** | **Description** | **Suggested relation to writing ability**[1] |
|---|---|---|---|
| WORD | | | |
| **Lexical complexity** | word length | number of letters per word (average) | + increases with ability |
| | word complexity | number of morphemes per word (average) | + increases with ability |
| | word frequency (50%) | proportion of words that overlap with 50% most frequent words of word frequency list | - decreases with ability |
| **Lexical richness** | adverbials | number of adverbials per clause (average) | + increases with ability |
| | type/token-ratio | number of different instances (tokens) per total number of lemmas (types) (average) | - decreases with ability |
| | prepositions | number of prepositions (per 1000 words) | + increases with ability |
| SENTENCE | | | |
| **Sentence complexity** | subordinate clauses | number of subordinate clauses (per 1000 words) | + increases with ability |
| | dependency length | distance between sentence components (average) | + increases with ability |
| | clauses | number of clauses (per 1000 words) | - decreases with ability |
| | sentence length | number of words per sentence (average) | + increases with ability |
| TEXT | | | |
| **Coherence** | argument overlap (lemmabuffer) | overlap between (lemmatized) arguments in preceding 10 words (per 1000 words) | - decreases with ability |
| | referential pronouns | number of 3rd person personal/possessive pronouns and demonstrative pronouns (per 1000 words) | - decreases with ability |
| | connectives | number connectives (of different categories) (per 1000 words) | - decreases with ability |

Analyses

The outcomes for the specific population in this study (i.e., primary school, grades 3 through 6) were examined in three ways for each measure presented in Table 13. First, for measures comprised of a specific set of words (i.e., prepositions, referential pronouns and connectives), all occurrences per word were mapped, and the distribution across grade level was evaluated. Because the number of pupils differed per grade (cf. Table 12) and essay length increased with grade level (cf. Figure 3), occurrences were computed per 100 essays of 60 words (i.e., the average essay length across grades).

Second, potential negative influences on reliability and validity were specified per measure, as well as their possible solutions. To detect potential pitfalls, essays eliciting extreme values (outliers) were analysed in order to identify the specific properties that caused these extremely high or low values. Outliers were specified as values beyond two standard deviations from the mean.

Finally, a developmental scale was constructed, based on the results of the selected measures. To select the exemplar essays provided on the scale, values for the separate measures presented in Table 13 were used to identify essays that illustrated different levels of text complexity. Three aspects of writing ability were defined: lexical sophistication (comprised of lexical complexity and lexical richness), sentence complexity, and coherence. For each of these three aspects, four exemplars were selected to compose a developmental scale: two essays per grade (i.e., grade 3 and grade 6), represented both low and high quality essays within the specific grade. The quality of the essays was based on human evaluation (cf. Section 4.2.3). Low quality essays and high quality essays were categorized in score group 1 and score group 4, respectively.

### 4.3.4 Results

**Composite measures: numbers and types of cases**

Prepositions

Appendix D lists all occurrences of prepositions within the current data set. Figure 4 presents the numbers of prepositions for which 10 or more occurrences were found within the data set. Figure 5 presents the density of these prepositions, that is, the number of prepositions based on a total of 60 words per essay (the average essay length across grades). Figure 6 presents the number of occurrences per specific preposition. These results showed that the overall increase in preposition use across grades (Figure 4 and Figure 5) was reflected in the occurrences per type of preposition (Figure 6). Almost all prepositions showed an increase across grades, although the differences in occurrence appeared small in some prepositions (e.g., *met*: with/by) and larger in others (e.g. *aan*: to/on/at).

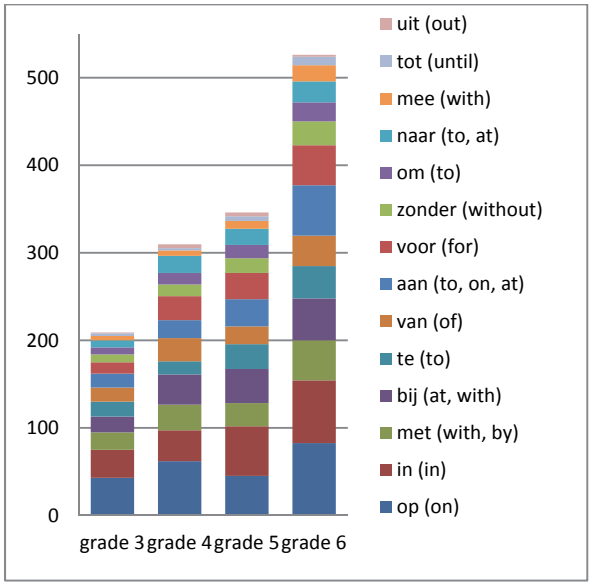



*Figure 4*. Prepositions: number of occurrences per grade level.

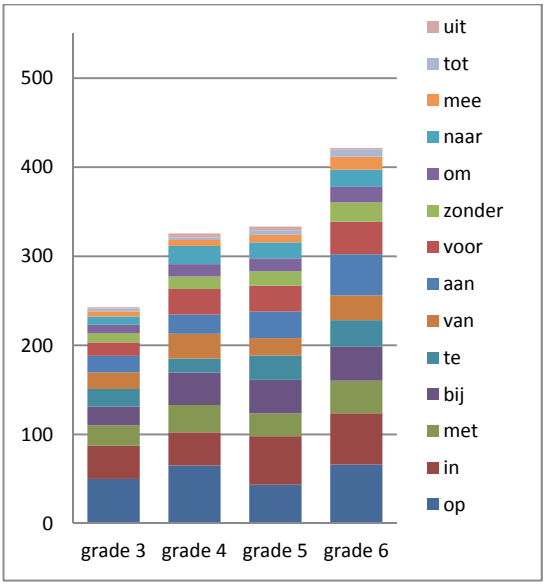



*Figure 5*. Prepositions: density* per grade level.

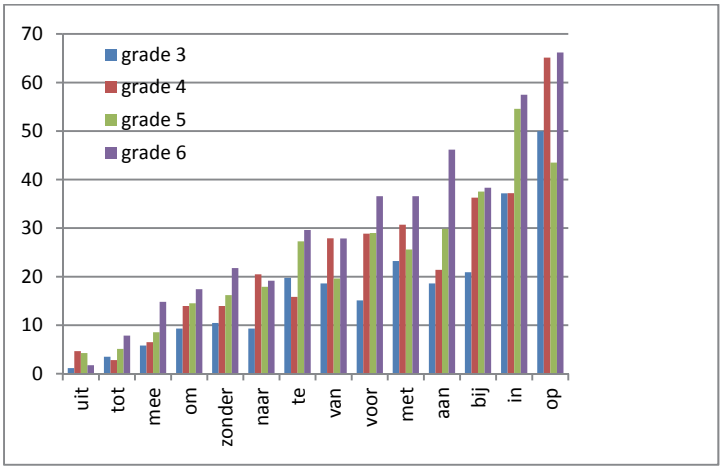



*Figure 6.* Prepositions: density* per word.

* N per 100 essays (60 words per essay)

Referential pronouns

In T-Scan, referential pronouns are defined as third-person personal and possessive pronouns (except "one") and demonstrative pronouns. Appendix D lists all occurrences of referential pronouns within the data set. Figure 7 presents the different referential pronouns for which 10 cases or more occurred within the data set and the number of occurrences per grade. Figure 8 presents the density per grade of these pronouns, and Figure 9 shows the density per specific pronoun. These results showed that on the one hand, the density of overall referential pronoun use per essay decreased with grade level (Figure 8); on the other hand, the separate types of pronouns followed different developmental patterns (Figure 9). For example, the use of the demonstrative pronouns *die* (that/which/who) and *dat* (that/which/what) was fairly stable across grades, whereas the use of the demonstrative *deze* (this/these) appeared to increase with grade level. The use of personal pronouns *hij* (he) and *hem* (him) on the other hand, seemed to decrease with grade level.

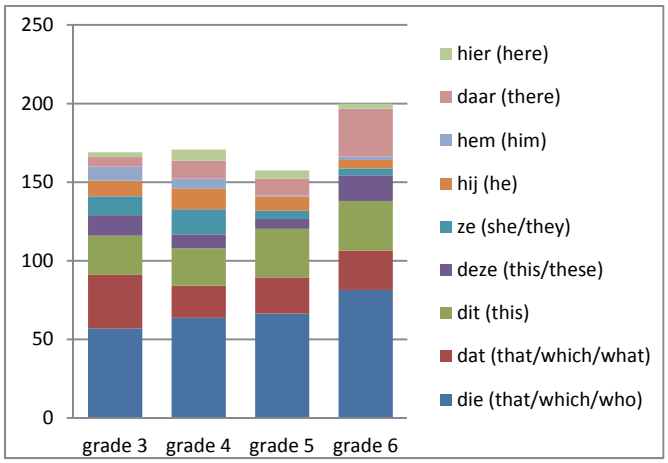



*Figure 7.* Referential pronouns: number of occurrences per grade level.

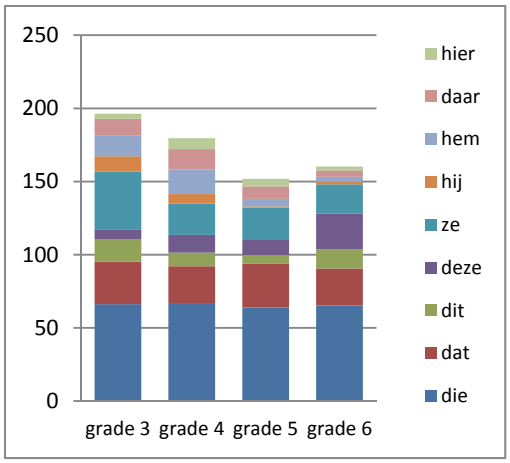



*Figure 8.* Referential pronouns: density* per word.

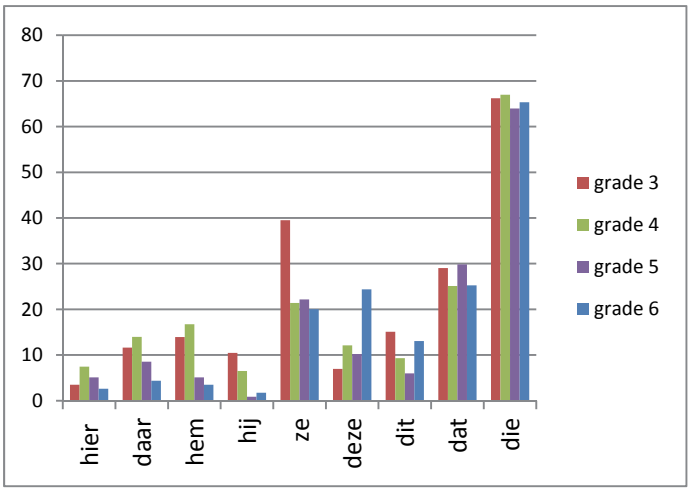



*Figure 9.* Referential pronouns: density* per grade level.

* N per 100 essays (60 words per essay)

Connectives

T-Scan uses a broad definition of connectives: in addition to words that connect two sentences, clauses or noun phrases (e.g. and, because, but) the list includes adverbs indicating time (e.g., already) or comparison (e.g. like). All connectives are grouped into five categories that indicate different types of relations (i.e., temporal, comparative, causal, enumerative and contrastive relations). Appendix D provides an overview of all connectives used within the data set with their frequencies. Figure 10 through Figure 21 present the absolute and relative numbers of occurrences per grade level for connectives that occurred in 10 items or more within the data set. Relative numbers (density) are based on 100 pupils per grade and 60 words per essay.

Figure 10 shows that the number of connectives used per text increased between grades. However, because the total number of words increased as well, the density of connectives occurring in essays decreased with grade level. Figure 11 presents the numbers of connectives per grade in an average essay length of 60 words. This figure indicates that the specific developmental pattern differed per category of connectives. The patterns per category are presented in Figure 12 through Figure 21.

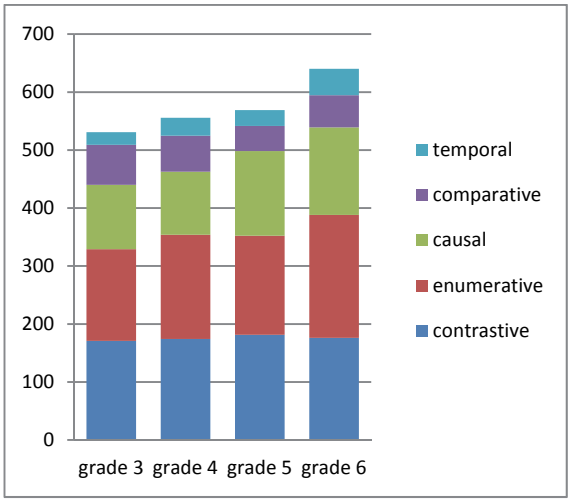



*Figure 10.* Connectives: number of occurrences per grade level.

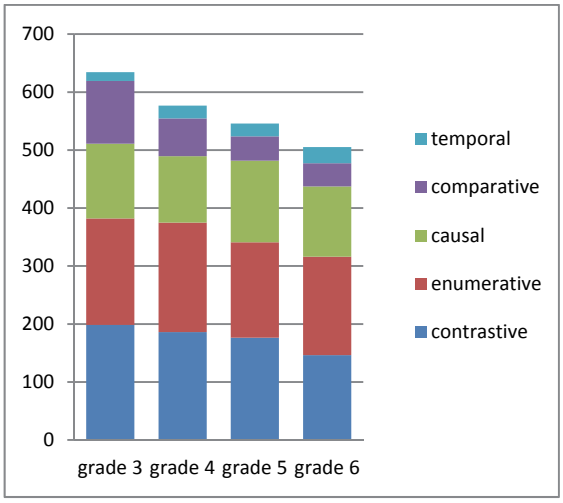



*Figure 11.* Connectives: density* per grade level.

* N per 100 essays (60 words per essay)

Because of the small differences in the use of both **comparative** and **temporal** connectives, significance could not be tested in these categories (cf. Table 7). However, the use of temporal connectives seemed to increase with grade level, a pattern mainly caused by the use of *toen* (then) because other temporal connectives did not seem to vary greatly among grades. The results for the group of comparative connectives are largely based on occurrences of *dan* (then), which showed a relatively large decline across grades, whereas *als* (if) showed a smaller—but seemingly steady—decrease.

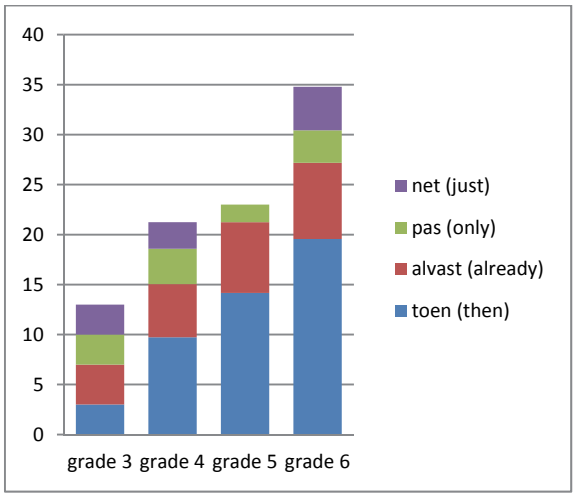



*Figure 12.* Temporal connectives: number of occurrences per grade level.

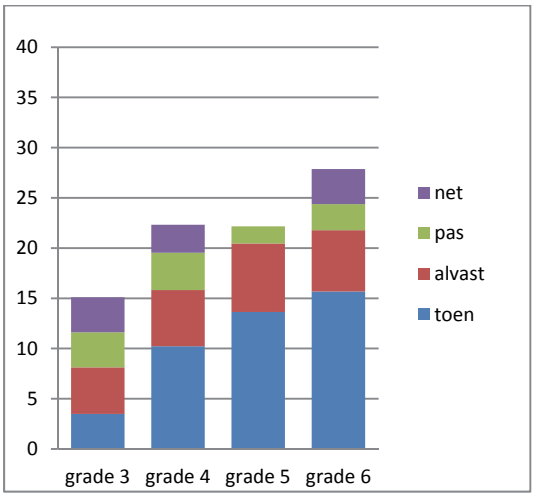



*Figure 13.* Temporal connectives: density* per grade level.

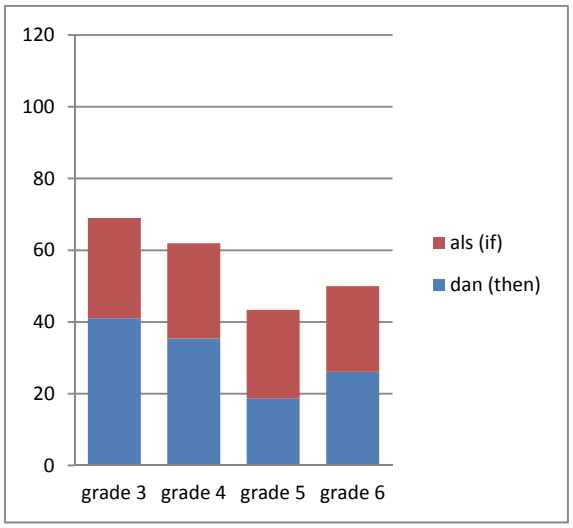



*Figure 14.* Comparative connectives: number of occurrences per grade level.

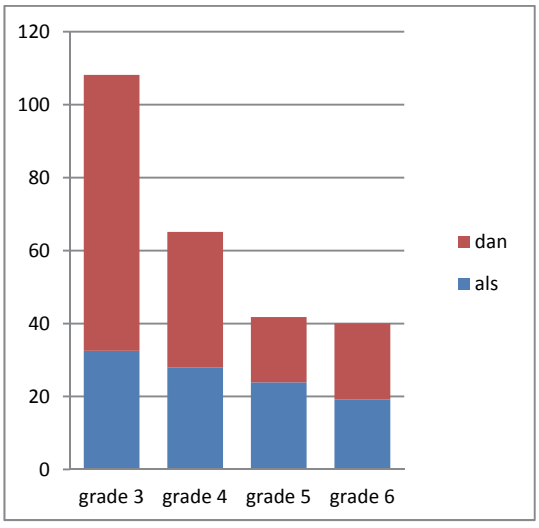



*Figure 15.* Comparative connectives: density* per grade level.

* N per 100 essays (60 words per essay)

The density of both **contrastive** and **enumerative** connectives was found to decrease significantly between grade 3 and grade 6 (cf. Table 7). In both categories, absolute numbers increased with grade level, which indicates that the decrease in density was caused by an increase in the number of other words, instead of a decrease in the number of connectives. Both groups were dominated by a single connective, for which a decline in (relative) numbers was found. Regarding enumerative connectives, the decrease in density seemed largely caused by a modest decrease in the density of *en* (and), whereas the group of contrastive connectives was dominated by *maar* (but), which showed a large decline in density across grades.

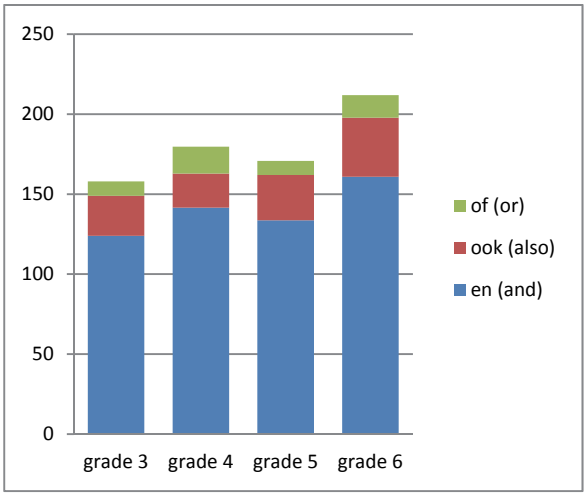



*Figure 16.* Enumerative connectives: number of occurrences per grade level.

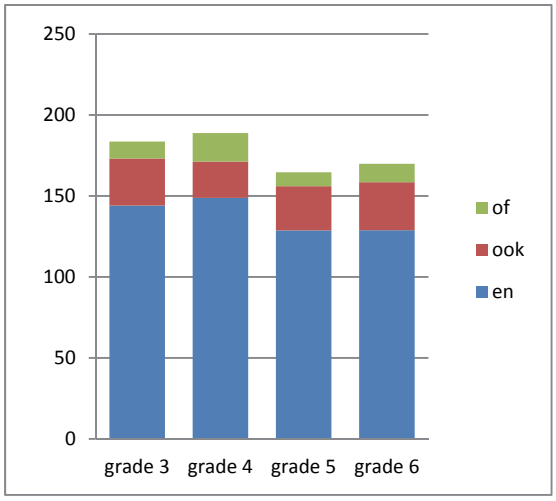



*Figure 17.* Enumerative connectives: density* per grade level.

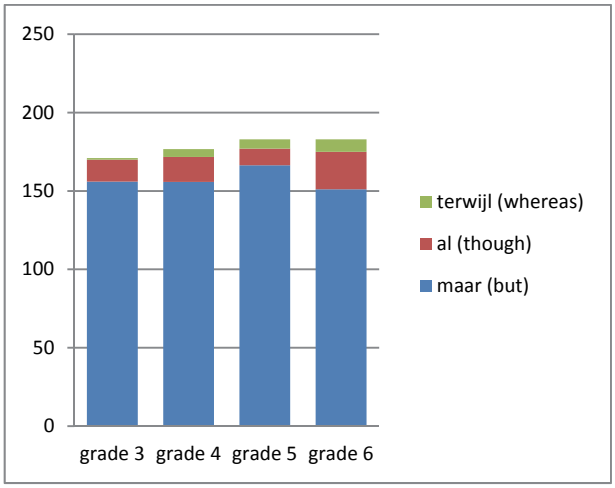



*Figure 18.* Contrastive connectives: number of occurrences per grade level.

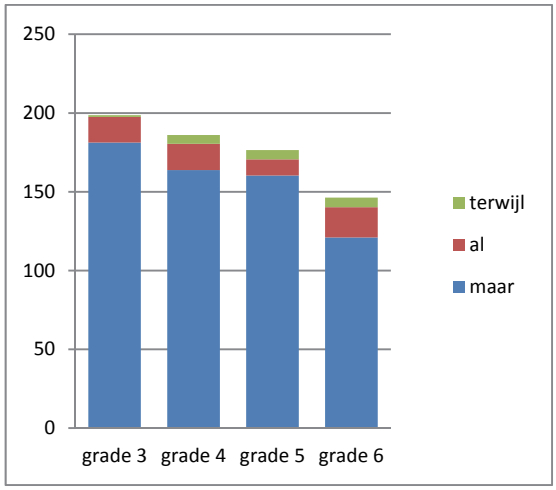



*Figure 19.* Contrastive connectives: density* per grade level.

* N per 100 essays (60 words per essay)

Finally, no clear developmental pattern could be determined for **causal connectives** (e.g., because, therefore, so), which reflects that no significant difference was found between grades 3 and 8 for this group of connectives (cf. Section 4.3).

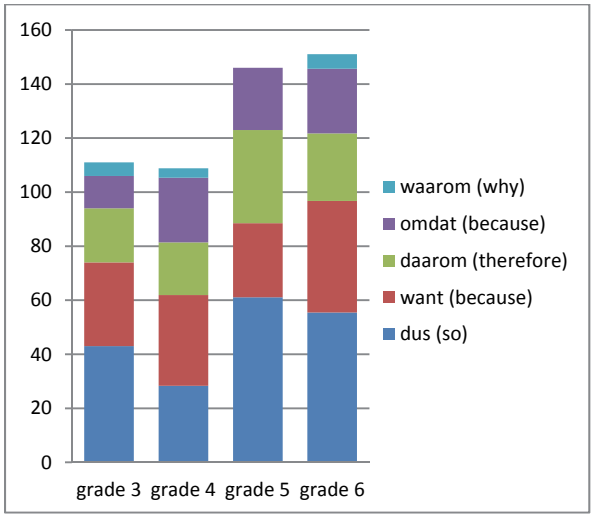



*Figure 20.* Causal connectives: number of occurrences per grade level.

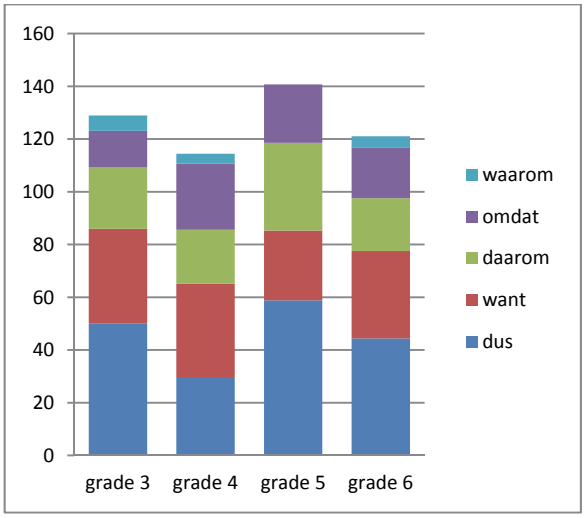



*Figure 21.* Causal connectives: density* per grade level.

* N per 100 essays (60 words per essay)

**Analysis of extreme cases**

In order to identify the possible factors that unjustly influenced the outcomes of T-Scan's measures and hence negatively influenced their validity, the specific features underlying T-Scan's measures of text complexity were analysed in essays that were identified as extreme cases, or outliers. An outlier was defined as an essay for which the value lies beyond standard deviations from the mean.

Table 14 through Table 16 presents the results of the analyses of the measures on sentence, word and text level. For each of the selected T-Scan measures, the mean value is given, in addition to an example of an essay that elicited both an extremely *low* value and an extremely *high* value. Based on the analysis of all outliers per measure, the possible sources of the extreme values were identified, and hypotheses regarding all factors that unjustly influenced the measures of text complexity were proposed.

Table 14. *Analysis of Outliers on Sentence Level*

| Measures of text complexity | Possible source of extreme text complexity score | Low complexity score | High complexity score |
|---|---|---|---|
| **Subordinate clauses**<br><br>*mean: 22.908* | text length: a short text induces a relatively high density of subordinate clauses<br>punctuation style: the use of periods instead of commas lowers the amount of subordinate clauses while the (complex) relation is still present | 0^^<br>ik heb niet genoeg punten.<br>Kan ik dan een telefoon krijgen?<br>want ik heb er maar acht.<br>ik wil een telefoon daarom heb ik het gedaan.<br>dat was niet zo slim.<br>maar ik een telefoon nodig!<br>dus ik hoop het maar. | 111.11<br>ik noem je maat omdat je mijn vriend bent. Als je mij een gratis telefoon geeft. en als je ze geeft dan gaan we een feestje bouwen. |
| **Dependency length**<br><br>*mean: 1.589* | sentence length: extremely short sentences lower dependency length<br>use of colloquial language: sentences resembling spoken language (flow of words) raise dependency length | 0.368<br>ik heb tien munten gespaard voor de actie. mijn naam is elske schocaten. dit is mijn telefoon nummer 06-22667757.<br>dit is mijn adres. en de munten zitten in de envelop. | 5.128<br>Ik heb maar 8 punten dus<br>ik doe er nog 2 wikkels bij want je kunt geen punten meer sparen want ze zijn uitverkocht dus ik kan niet meer sparen. |
| **Clauses^**<br><br>*mean: 155.595* | use of colloquial language: the density of clauses is based on the number of finite verbs; sentences resembling spoken language raise clause density | 239.13<br>Ik heb geen 10 spaarpunten maar ja ik probeer het. Kan ik dan alsjeblieft de Telefoon krijgen. Een week later ik heb 'm ik heb 'm hij is nep. Er staat op je krijgt een echte. 1 jaar later ik heb 'm en hij is echt. | 87.72<br>ik kan geen 10 punten krijgen want bij de repen chocola staan geen punten meer. en ik hoef er nog maar twee.<br>en als ik geen telefoon krijg koop ik niks meer van smikkel. en mijn hele familie en vriends familie en mijn vriends vriends ook niet lekker puh. |
| **Sentence length**<br><br>*mean: 11.616* | use of colloquial language: sentences resembling spoken language (flow of words) increase sentence length | 4.600^^<br>Ik heb acht stickers.<br>Maar geen tien stickers.<br>En ik wil die telefoon.<br>Mag ik die telefoon asjeblieft.<br>Dan krijgen jullie 4,50 ook. | 47.00<br>Ik heb gestuurd 8 punten en twee smikkels (zonder punten) en daarover wil ik praten: Want ik zie nergens meer smikkels (met punten) en ik kan er niks aan doen dat ik die twee punten niet heb, maar ik zou heel graag toch dat telefoontje willen krijgen. |

^ reverse complexity measure: low score indicates high complexity
^^no true outlier: within 2 sd from mean

Table 15. *Analysis of Outliers on Word Level*

| Measures of text complexity | Possible source of extreme text complexity score | Low complexity score | High complexity score |
|---|---|---|---|
| **Word length**<br><br>*mean: 3.983* | task: frequent use of (long) task-dependent words influences average word length | 3.250<br>Ik heb geen tien punten ik heb er maar acht. maar ik wil heel graag die telefoon. Ik hoop dat ik hem win. Dat zal echt leuk zijn. Denk ik ik hoop dat ik hem win. Ik hoop het echt. Maar ik heb maar acht punten en het is 3 april en het is tot en met 30 april en ik wil 10 punten. | 5.227<br>Ik heb net een nieuwe lekkere chocoladereep uitgevonden en ik wil graag een nieuwe telefoon hebben. Wilt U dan twee chocoladerepen ruilen? |
| **Word complexity**<br><br>*mean: 1.223* | task: frequent use of (complex) task-dependent words influences word complexity | 0.191<br>in de winkel zijn geen punten meer maar ik wil er zo graag een. het was niet eens 30 april maar ik had er maar 8. maar ik heb geen telefoon en ik wil er zo graag een. | 0.311<br>Ik heb net een nieuwe lekkere chocoladereep uitgevonden en ik wil graag een nieuwe telefoon hebben. Wilt U dan twee chocoladerepen ruilen? |
| **Word frequency^**<br><br>*mean: 0.551* | text length: influence of 'infrequent' words (according to frequency list) is relatively large when text is short<br><br>* S.v.p. : s'íl vous plait<br>**A.u.b: alstublieft<br>(common abbreviations) | 0.737<br>in de winkel zijn geen punten meer maar ik wil er zo graag een. het was niet eens 30 april maar ik had er maar 8. maar ik heb geen telefoon en ik wil er zo graag een. | 0.326<br>Ik heb maar 8 punten maar ik wil die telefoon zo graag hebben. Alstublieft.<br>Firma Smikel krijgt 400 euro van mij. Maar 8 repen en 8 punten. Alstublieft.<br>Hoeveel wilt u dan? 800-600-300-100-200?<br>Wil u een scooter?<br>S.v.p.*<br>A.u.b.**<br>S.v.p.<br>A.u.b.<br>S.v.p. |
| **Adverbials**<br><br>*mean: 0.954* | text length: a short text induces a relatively large density of adverbials | 0<br>ik heb 10 punten.<br>en mag ik een telefoon.<br>en u krijgt 2.50. dus ik stuur een brief en jullie geven mij een telefoon. | 2.400<br>Er was nergens in de winkels nog de twee laatste punten te vinden. het was dus onmogelijk om aan de tien repen te komen.<br>Ik heb daarom de acht punten plus twee hele winkels naar u opgestuurd omdat ik toch graag de telefoon wil ontvangen.<br>Ik kan er niets aan doen dat het er maar acht zijn. |
| **Type/token-ratio^ (lexical diversity)**<br><br>*mean: 0.699* | text length: a short text reduces the chance of several tokens per type, hence increasing type/token-ratio | 0.955<br>Ik kan helaas geen 10 spaarmunten krijgen. Want de winkels geven ze niet meer uit en ik wil wel graag er een. | 0.422^^<br>bij de winkel zijn geen repen meer met punten. dus ik wil ze opsturen met 2 zonder punten erbij maar het is nog geen 30 april maar die 2 repen die zijn wel heel hoor.<br>maar ik vind het wel jammer dat er geen repen meer zijn met geen punten.<br>maar nou heb ik wel 8 punt maar dan 2 repen erbij zonder punten.<br>maar ik kan er echt niets aan doen.<br>maar ik wil de telefoon heel graag ontvangen.<br>ik heb maar 8 punten dus ik doe er maar 2 hele repen bij zonder punten<br>maar ik vind het heel erg. |
| **Prepositions**<br><br>*mean: 61.943* | text length: a short text induces a relatively high density of prepositions | 0^^<br>Ik had geen 10 munten want de winkels hadden geen chocola meer en daarom had ik geen tien munten. Ik had er maar acht en dat is niet genoeg want je had er tien nodig. Maar je had wel een kans en dat is goed maar je hebt je kans niet verknald. | 157.9<br>De punten van smikkel zitten er niet meer op dus jullie krijgen acht punten en twee repen zonder punten. |

^ reverse relation to writing ability: low score indicates high ability
^^ no true outlier: within 2 sd from mean

Table 16. *Analysis of Outliers on Text Level*

| Measures of text complexity | Possible source of extreme text complexity score | Low complexity score | High complexity score |
|---|---|---|---|
| **Argument overlap**<br><br>*mean: 78.371* | text length: short texts in which a number of words are repeated elicit a high argument overlap | 0^^<br>Ik spaar ook mee met de actie voor een gratis telefoon.<br>Maar ik heb maar 8 spaarpunten kunnen vinden Op de wikkels van de repen staan geen spaarpunten meer. Daarom heb ik er maar 2 wikkels bij gedaan waar geen spaarpunten op staan. Ik hoop dat u daarmee akkoord gaat.<br>Want in geen ene supermarkt of andere winkels staan er nog spaarpunten meer op de wikkels.<br><br>Bij voorbaat dank. | 234.375<br>Ik heb geen tien punten ik heb er maar acht.<br>maar ik wil heel graag die telefoon.<br>Ik hoop dat ik hem win.<br>Dat zal echt leuk zijn.<br>Denk ik. Ik hoop dat ik hem win.<br>Ik hoop het echt.<br>Maar ik heb maar acht punten en het is 3 april en het is tot en met 30 april<br>En ik wil 10 punten. |
| **Referential pronouns**<br><br>*mean: 31.871* | coherence vs. cohesion: relations between clauses can also be expressed implicitly | 0^^<br>ik heb maar 8 punten,<br>ik wil graag de telefoon hebben.<br>maar de actie is voorbij.<br>kan ik ook gewone wikkels geven.<br>om de telefoon te krijgen. | 145.16<br>Ik heb acht punten. Maar ik kan niet meer doorsparen, omdat de winkels die die reep verkopen, die hebben geen punten meer daardoor kan ik niet meer doorsparen. Want ik wil die telefoon heel graag kan u hem alsnog opsturen, want ik wil hem zo graag die telefoon. Alstublieft kan die telefoon opsturen. Als u dat wilt dan heel erg bedankt. |
| **Causal connective**<br><br>*mean: 35.522* | correctness: wrong use of connective is counted as connective<br>text length: short letter with one or two connectives elicits a relatively high value<br>task: not all text goals elicit (explicit) causal relations | 0^^<br>in de winkel zijn geen punten meer maar ik wil er zo graag een.<br>het was niet eens 30 april maar ik had er maar 8. maar ik heb geen telefoon en ik wil er zo graag een. | 68.19<br>Ik heb een gek iets over de SMIKKELrepen. Er staan geen spaarpunten op de wikkels. Daarom kan ik de telefoon niet winnen. Daarom heb ik twee wikkels zonder spaarpunten. Dus eigenlijk heb ik wel de telefoon gewonnen. in de brief zit 2,50 voor postzegels. |
| **Comparative connective**<br><br>*mean: 0.546* | alternative meaning: homonyms of connective are sometimes counted as connective, e.g. *vindt u het goed als…* [Do you approve if …]<br>text length: a short text with one or two connectives elicits a relatively high value<br>task: not all text goals elicit (explicit) comparative relations | 0^^<br>het spijt me dat ik geen 10 punten bij elkaar heb verzameld. ik heb erg mijn best gedaan maar het is niet gelukt. ik wou zo graag die telefoon hebben. mijn geld heb ik uitgegeven aan die repen. en ik moest nog 2 punten, maar ik vond ze nergens. ik heb overal gezocht, maar nergens kon ik er een meer vinden en daarom heb ik die 2 punten laten zitten. maar dat is toch niet eerlijk? ik heb niet nagedacht. maar ik kan er niks aan doen dat ik geen 10 punten heb. ik had allemaal gezocht maar geen punten. er waren geen punten meer. ik hoop dat u mij begrijpt. | 29.41<br>Ik koop vaak smikkels. Maar ik graag die telefoon.<br>Daarvoor moet je 10 punten hebben.<br>Maar ik heb er maar acht.<br>Vindt u het goed als ik 8 punten en 2 wikkels erbij doe? |
| **Contrastive connective**<br><br>*mean: 7.576* | alternative meaning: homonyms of connective are counted as connective, e.g. 'maar' is a connective as well as an adverb: *ik hoop het maar [I do hope so]; ik heb er maar acht [I have only 8]*<br>text length: short letter with one or two connectives elicits a relatively high value<br>task: not all text goals elicit (explicit) contrastive relations | 0^^<br>ik had m'n tien punten meegenomen voor uit te sloven en de pestkop van de school had er 8. ik had ze per ongeluk laten liggen. en hij was aan het trefballen en hij bracht de bal binnen en heeft twee punten gejat. dus nu heb ik er acht.<br><br>mag ik alstublieft de telefoon. | 76.92<br>Ik heb ondertussen 5 wikkels gehad maar ondertussen zat er op geen een een munt. Ik stuur dit op omdat ik graag de telefoon wil ontvangen. |

| | | | |
|---|---|---|---|
| **Enumerative connective**<br><br>*mean: 30.910* | text length: short letter with one or two connectives elicits relatively high value<br>coherence vs. cohesion: relations between clause scan also be expressed implicitly | 0^^<br>bij de winkel zijn geen repen meer met punten. dus ik wil ze opsturen met 2 zonder punten erbij maar het is nog geen 30 april maar die 2 repen die zijn wel heel hoor.<br>maar ik vind het wel jammer dat er geen repen meer zijn met geen punten.<br>maar nou heb ik wel 8 punt maar dan 2 repen erbij zonder punten.<br>maar ik kan er echt niets aan doen.<br>maar ik wil de telefoon heel graag ontvangen.<br><br>ik heb maar 8 punten dus ik doe er maar 2 hele repen bij zonder punten maar ik vind het heel erg. | 125.00<br>ik heb 10 punten. en mag ik een telefoon. en u krijgt 2.50. dus ik stuur een brief en jullie geven mij een telefoon. |
| **Temporal connective**<br><br>*mean: 5.455* | text length: short letter with one or two connectives elicits relatively high value<br>task: not all text goals elicit (explicit) comparative relations | 0^^<br>Ik kan helaas geen 10 spaarmunten krijgen. Want de winkels geven ze niet meer uit en ik wil wel graag er een. | 60.24<br>Ik heb maar acht punten, maar het is al 20 april. Maar ik eet altijd smikkelchocola. Maar ik ben drie weken op vakantie geweest en ik kwam gisteren terug dus, en ik las vandaag pas de actie. Maar ik wil zo graag die telefoon want mijn andere telefoon is kapot. Heb a.u.b. medelijden, ik wil die telefoon zo graag mag ik alstublieft die telefoon.<br>Als u mij een brief wil sturen stuur dan naar: Izaak Den Dekker. alvast bedankt en tot ziens |

^^ no true outlier: within 2 sd from mean

Alternative analyses per clause

The results on the analysis of extreme cases presented in Table 14 through Table 16 indicated that text length influences several text complexity scores. Therefore, an additional analysis was performed, in which several T-Scan measures were converted. Instead of computing the number of instances per 1000 words, the average values per clause were calculated for prepositions, referential pronouns and connectives. This conversion was expected to reduce the "overvaluation" of the use of these words in very short texts (cf. Table 16, example of enumerative connectives). Figure 22 through Figure 29 present the results of this additional analysis.

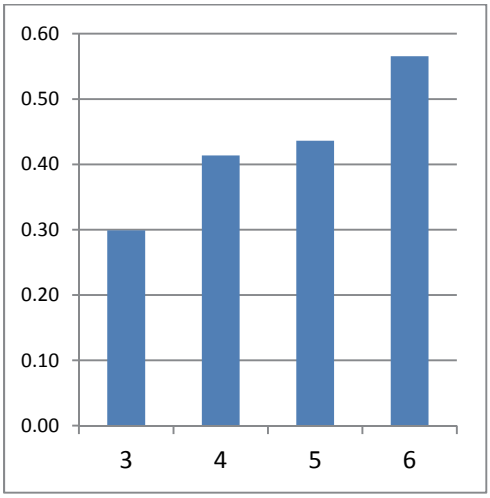



*Figure 22.* Prepositions: average number per clause.

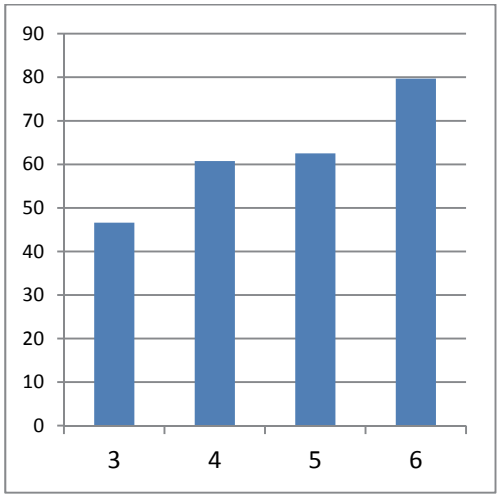



*Figure 23.* Prepositions: average number per 1000 words.

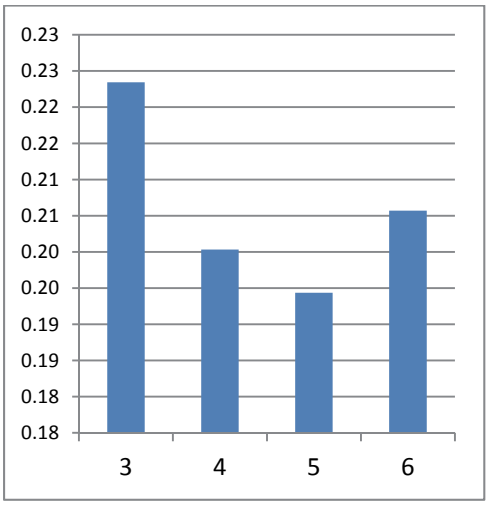



*Figure 24.* Referential pronouns: average number per clause.

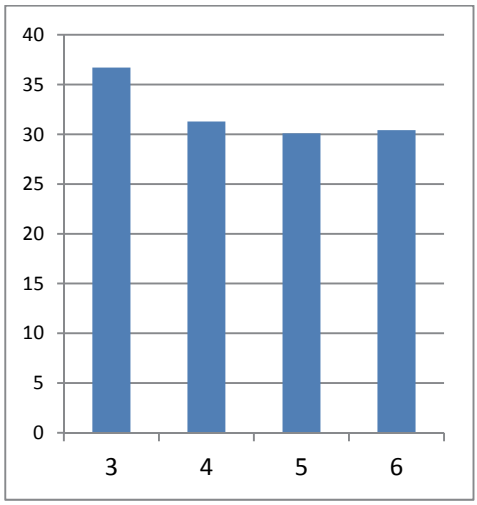



*Figure 25.* Referential pronouns: average number per 1000 words.

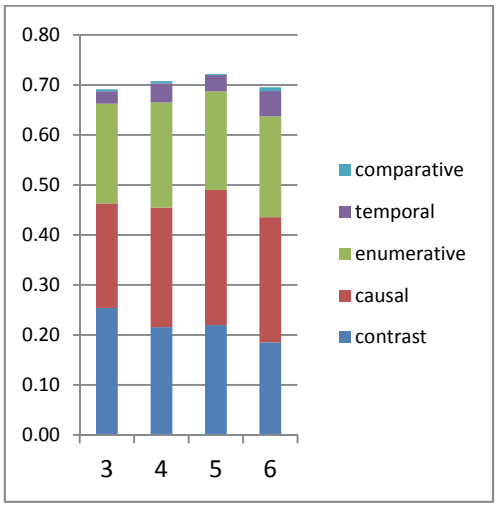



*Figure 26.* Connectives: average number per clause.

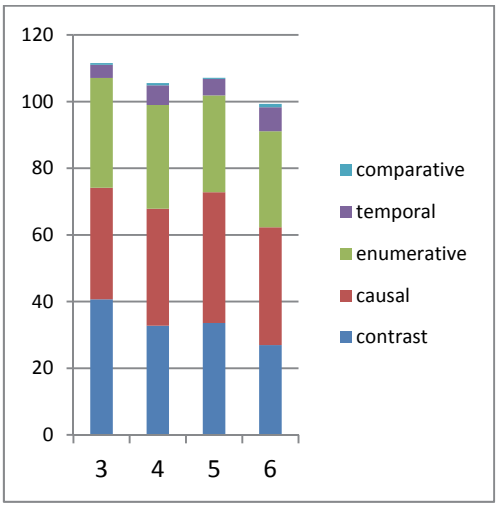



*Figure 27.* Connectives: average number per 1000 words.

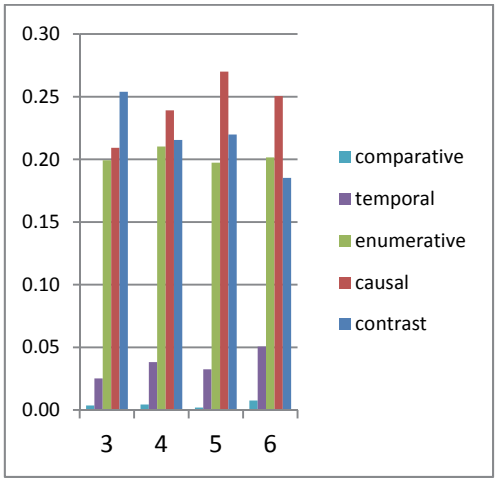



*Figure 28.* Connectives: average number per clause.

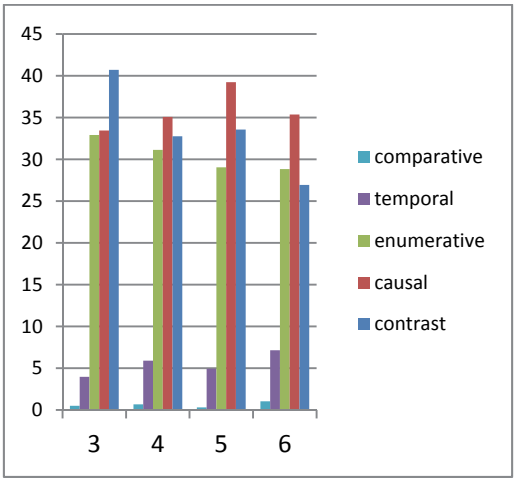



*Figure 29.* Connectives: average number per 1000 words.

For prepositions, the results per clause (Figure 22) demonstrated a similar relation to writing ability, compared to the results of the original density measures (Figure 23), that is, the number of prepositions per clause increased with grade. The overall pattern of referential pronouns seemed unaffected by the conversion (Figure 24 and Figure 25). However, the differences between grades are more prominent when the number of referential pronouns was computed per clause. For connectives, the slight decrease in overall connective use was flattened by the conversion from connective density to the average number of connectives per clause (Figure 26 and Figure 27). For individual clusters of connectives, however, there appeared to be no differences in their relation to writing ability (Figure 28 and Figure 29). The above results are reflected in the correlations between average numbers per clause and per 1000 words, as shown in Table 17.

Table 17. *Correlations between Measures Based on Averages per Clause and per 1000 Words*

| Measure | Correlation |
|---|---|
| prepositions | .96 |
| referential pronouns | .96 |
| connectives | .89 |

To evaluate further the effect of the conversion from density measures to clause-based measures, the ordering of essays based on the values of each measure (i.e., prepositions, referential pronouns, and connectives) was compared according to both methods. In addition to the order of the individual essays, the 10 highest-scoring essays were considered in order to examine the effect on these extreme values. The 10 lowest-scoring essays were not considered because all essays lacking the features under consideration received a value of 0 and could therefore not be considered "extreme" values. Comparison of the ordering of essays based on values per clause as well as values per 1000 words, the individual ordering was found to change, but little effect on the outliers was found. That is, little change was found in the group of highest-scoring essays when the values per clause were computed, indicating that the effect of text length was not yet eliminated. Table 18 through Table 20 illustrate this finding.

Table 18. *Prepositions: Text Length of 10 Highest-scoring Essays*

| | Type of measure for selection | |
|---|---|---|
| | per clause | per 1000 words |
| av. number of words | 61.1 | 42.4 |
| minimum | 19.0 | 19.0 |
| maximum | 126.0 | 66.0 |
| standard deviation | 26.6 | 16.5 |

Table 19. *Referential Pronouns: Text Length of 10 Highest-scoring Essays*

| | Type of measure for selection | |
|---|---|---|
| | per clause | per 1000 words |
| av. number of words | 46.7 | 43.3 |
| minimum | 26.0 | 26.0 |
| maximum | 77.0 | 62.0 |
| standard deviation | 14.3 | 10.0 |

Table 20. *Connectives: Text Length of 10 Highest-scoring Essays*

| | Type of measure for selection | |
|---|---|---|
| | per clause | per 1000 words |
| av. number of words | 55.5 | 60.4 |
| minimum | 26.0 | 26.0 |
| maximum | 104.0 | 104.0 |
| standard deviation | 20.0 | 22.7 |

The conversion from density measures to averages per clause did not seem to limit the effect of text length greatly. Taking into account that text length and clause length are highly related (r = .92), the solution to dealing with differences in text length is likely found in methods used to standardize text length, instead of a shift to clause-based measures. However, the use of clauses instead of sentences as units of measurement does have the advantage of being independent of punctuation because the identification of clauses is based on the presence of finite verbs, not sentence stops. Hence, the conversion from density measures to averages per clause might eliminate the need to correct punctuation in order to elicit valid results.

**Developmental scales**

To evaluate the interpretability of text complexity scores as measures of writing ability, developmental scales with exemplar essays were composed for the selected measures of word, sentence and text level (cf. Table 13). To select the exemplars, the essays were first grouped by human essay score (cf. Table 6). Essays of "low achievers" (i.e., score group 1) and "high achievers" (i.e., score group 4) representing pupils in the middle year and final year of primary education were selected, based on their complexity scores for word, sentence and text level measures.

The results are presented in Figure 30 through Figure 32. In these figures, the development of writing is illustrated by means of exemplars written by pupils differing in grade (3 or 5) and ability level (below or above average). Hence, the figures provide input for an across-grade developmental comparison when they are read horizontally, and a within-grade normative comparison of text complexity when they are read vertically (cf. Dean and Quinlan, 2010). Specifically, the developmental scales provide insight into the relation between text complexity (i.e., feature values in T-Scan), stage of writing development (i.e., grade), and writing proficiency (i.e., low/high achiever). By offering a comparison between exemplars, these developmental scales provide information about the degree to which, according to human judgment, text complexity reflects the development in writing ability and the degree to which specific features of complexity can help us understand our perceptions of the quality of a writing performance.

| | | |
|---|---|---|
| **Level: middle year of primary education (grade 3, age 8/9)** | | |
| | *low achiever* | *high achiever* |
| **Lexical sophistication** | ik heb 8 punten gespaard.<br>maar nu kan ik geen repen met punten<br>en het is geen 30 april. Dus nu heb ik geen punten en ik wil wel een telefoon.<br>ik heet leanne ik woon in ameide. | Ik kan geen tien munten opsturen, want er waren geen wikkels met spaarmunten meer. Daarom heb ik 2 alleen maar wikkels erbij gedaan, en acht spaarmunten.<br>Ik wil wel graag die telefoon krijgen, als het niet mogelijk is dat ik die telefoon toch krijg dan hoor ik dat graag. Ik wil best nog €5,50 opsturen?<br>Want ik wil het gewoon super graag.<br>Want mijn mobiel is heel erg stuk.<br>Stuur je hem op als het kan en laat het horen als het niet kan. |
| word length | 3.53 | 4.06 |
| word complexity | 1.18 | 1.19 |
| word frequency | 0.61 | 0.54 |
| type/token ratio | 0.68 | 0.61 |
| adverbial clauses | 0.57 | 1.14 |
| prepositions | 52.63 | 23.53 |
| **Level: final year of primary education (grade 6, age 11/12)** | | |
| | *low achiever* | *high achiever* |
| **Lexical sophistication** | Als je 1 smikkel koopt krijg je een spaarpunt. Als je er 10 hebt moet je het opsturen (dat staat beneden).<br>De actie is tot met 30 April.<br>Als je er 8 hebt en er zijn geen spaarpunten meer maar wel smikkels koop er 2 smikkels. en stuur ze op en je 8 punten.<br>dan krijg je telefoon nog steeds. maar dan wel voor 30 april. Stuur ook €2,50 voor de opsturen van de telefoon. en voor de spaarzegel. | Ik zag laatst de actie dat je een gratis telefoon kunt winnen, dat leek mij dus wel leuk om hieraan mee te doen!<br>Tot mijn verbazing zijn er nergens meer spaarpunten te vinden ondanks het nog geen dertig april is.<br>Ik heb in de envelop acht spaarpunten gedaan met twee hele wikkels zonder punten, maar eigenlijk verwacht ik nu wel die telefoon, want ik kan er niets aan doen dat er geen punten meer zijn… |
| word length | 3.72 | 4.29 |
| word complexity | 1.17 | 1.23 |
| word frequency | 0.58 | 0.53 |
| type/token ratio | 0.57 | 0.73 |
| adverbial clauses | 1.00 | 1.44 |
| prepositions | 111.11 | 133.33 |

*Figure 30.* A developmental scale of writing ability (exemplars evaluated on word level features).

| | Level: middle year of primary education (grade 3, age 8/9) | |
|---|---|---|
| | *low achiever* | *high achiever* |
| **Sentence complexity** | ik heb 10 punten.<br>en mag ik een telefoon.<br>en u krijgt 2.50. dus ik<br>stuur een brief en jullie geven<br>mij een telefoon. | Ik ben Valerie.<br>Ik heb 8 spaarpunten en 2 wikkels opgestuurd. Ik heb niet genoeg, en wil toch die telefoon hebben. Ik kon geen spaarzegels meer krijgen maar de actie is nog niet afgelopen. Willen jullie mij alsjeblieft de telefoon geven want ik kon ook niet weten dat de dingen niet verkrijgbaar meer zijn. Ik vind het maar raar. |
| subordinate clauses | 0 | 16.59 |
| dependency length | 1.53 | 1.89 |
| clauses | 208.33 | 203.39 |
| sentence length | 8.00 | 9.8 |
| | **Level: final year of primary education (grade 6, age 11/12)** | |
| | *low achiever* | *high achiever* |
| **Sentence complexity** | Als je 1 smikkel koopt krijg je een spaarpunt. Als je er 10 hebt moet je het op sturen (dat staat beneden).<br>De actie is tot met 30 april.<br>Als je er 8 hebt en er zijn geen spaarpunten meer maar wel smikkels koop er 2 smikkels. en stuur ze op en je 8 punten.<br>dan krijg je telefoon nog steeds. maar dan wel voor 30 april. Stuur ook €2,50 voor de opsturen van de telefoon. en voor de spaarzegel. | Ik stuur u deze brief omdat ik niet meer voor de telefoon kan sparen.<br>Ik ben hard aan het sparen voor die telefoon maar er zitten geen spaarmunten meer op de repen.<br>En het is nog geen 30 april!<br>Daarom heb ik 2 repen SMIKKEL erbij gedaan zodat ik toch nog aan 10 punten zit.<br>Ik zou heel blij zijn als ik toch nog die telefoon zou mogen ontvangen. |
| subordinate clauses | 37.04 | 42.86 |
| dependency length | 2.25 | 2.96 |
| clauses | 148.15 | 142.86 |
| sentence length | 11.57 | 14.00 |

*Figure 31.* A developmental scale of writing ability(exemplars evaluated on sentence level features).

| | Level: middle year of primary education (grade 3, age 8/9) | |
|---|---|---|
| | *low achiever* | *high achiever* |
| **Coherence** | Ik heb geen 10 punten maar ik heb heel hard gewerkt.<br>Ik vind jullie actie top<br>Ik wil graag 10 punten.<br>Stuur mij alstublieft die telefoon. | Ik ben Valerie.<br>Ik heb 8 spaarpunten en 2 wikkels opgestuurd. Ik heb niet genoeg, en wil toch die telefoon hebben. Ik kon geen spaarzegels meer krijgen maar de actie is nog niet afgelopen. Willen jullie mij alsjeblieft de telefoon geven want ik kon ook niet weten dat de dingen niet verkrijgbaar meer zijn. Ik vind het maar raar. |
| argument overlap | 117.65 | 84.75 |
| referential pronouns | 58.82 | 33.9 |
| connectives | 176.47 | 101.7 |
| | **Level: final year of primary education (grade 6, age 11/12)** | |
| | *low achiever* | *high achiever* |
| **Coherence** | Hallo mijn naam is junior en ik had een vraagje: in mijn winkel in de buurt is een smikkelvoorraad maar er is geen spaarpunt op. Dus ik wou vragen of ik er acht mocht sturen<br>met twee smikkels. Dank u. | Ik zag laatst de actie dat je een gratis telefoon kunt winnen, dat leek mij dus wel leuk om hieraan mee te doen!<br>Tot mijn verbazing zijn er nergens meer spaarpunten te vinden ondanks het nog geen dertig April is.<br>Ik heb in de envelop acht spaarpunten gedaan met twee hele wikkels zonder punten, maar eigenlijk verwacht ik nu wel die telefoon, want ik kan er niets aan doen dat er geen punten meer zijn. |
| argument overlap | 50.0 | 40.0 |
| referential pronouns | 50.0 | 26.67 |
| connectives | 100.0 | 53.33 |

*Figure 32.* A developmental scale of writing ability (exemplars evaluated on text level features).

### 4.3.5 Discussion

In this section, the validity of text complexity measures as indicators of writing ability is explored by performing a qualitative analysis of the measures used in T-Scan to determine text complexity. For a selection of text complexity measures for which a relation with writing ability was found (cf. Section 4.2), their validity as indicators of writing ability was evaluated. For this purpose, the measures' relation to writing was hypothesised based on the outcomes of agreement with both grade level and human essay scores. These hypotheses were then tested in three ways. First, for composite measures, the output from T-Scan was analysed, and the developmental patterns per individual component were illustrated. Furthermore, for all selected measures, the factors unjustly influencing text complexity scores were identified by analysing outliers per measure. Finally, the interpretability of the selected measures was illustrated by using a developmental scale of writing ability.

The analyses of the outcomes of the composite measures revealed differences among individual instances of referential pronouns, connectives and writing ability. These results indicated that, in the present population, not all instances of the measures demonstrated the same relation to writing ability. That is, the use of some words was related to a high writing ability (e.g., *terwijl*, whereas; *hij*, he), whereas the use of other words in the same category (e.g. *maar*; but *deze*, this/these) indicated low ability. These differences were likely caused by the fact that writing ability was still developing in the population sample.

For example, novice writers have not yet mastered all the words indicating coherence relations (i.e., connectives and referential pronouns). Consequently, some highly able pupils produced cohesive devices that indicated complex coherence relations, whereas others failed to express these relations correctly because they produced the wrong connective. Less skilled writers, on the other hand, sometimes used a relatively large number of cohesive elements to express a fairly simple coherent relation, or they use basic connectives incorrectly. The result of these developmental differences between pupils was that the usage patterns of these measures were difficult to predict, which complicates the interpretation of their relation to writing ability.

The use of prepositions was found to decrease with grade level, and all instances seem to indicate a similar relation to writing ability, which means no indication of different developmental patterns within the category of prepositions was found (Figure 4 through Figure 6). This finding may be because the prepositions used by the population did not differ greatly in their conceptual complexity and in Dutch, the syntactical complexity of prepositions does not vary much. Regarding referential pronouns, on the other hand, differences were found in individual cases (Figure 7 through Figure 9). In general, the use of referential pronouns was negatively correlated to writing ability. In many individual cases, however, no clear relation with grade level was found. In contrast, the use of the demonstrative *deze* (this/these) appeared to increase with grade level. Referring to previously stated entities, *deze* is more formal and common in written language, whereas in

spoken language, *die* is typically used. Furthermore, the use of all personal pronouns in the category, that is, *hij* (he), *hem* (him) and *ze* (she), appears to decrease with grade level. These findings indicated that as writing ability increased, writers chose to use distant and formal language more often and refer to actual persons less often.

With regard to connectives, a pattern was found in which the overall density of connectives decreased with grade level. This finding might indicate that proficient writers do not need to rely on (explicit) cohesive elements to achieve coherent relations within their texts. The absolute numbers of connectives, however, showed an increase, and the pattern of connective use differed per category. Several factors likely influenced the results of the coherence measures for novice writers. First, not all connectives reflected an equally level of ability because they differed in their conceptual and syntactical difficulty. Analyses of spoken language have shown that both the *types of relations* that are demonstrated by connectives and the *grammatical structure* they require influences the acquisition order of connectives (Evers-Vermeul & Sanders, 2009). This implies that, based on their relative complexity, different developmental patterns are demonstrated for different types of connectives, and that their (correct) use should be valued accordingly when assessing writing ability.

In T-Scan, however, all connectives contribute equally to the coherence score, which may lead to unexpected developmental patterns. For example, the use of *maar* (but) as a connector between sentences, instead of *en* (and) or a period, will cause a relatively high density of causal connectives, whereas it indicates a low level of writing ability (cf. Figure 33, example A). In Dutch, the causal *maar* is used to demonstrate a (negative) additive relation as well as a (negative) causal relation. In the first case, *maar* serves as a conjunct between two clauses (A and B), while in the second case, an implication can be deduced (A →B) and *maar* can be replaced by *hoewel* (although) (cf. Sanders, Spooren, & Noordman, 1992; Evers-Vermeul & Sanders, 2009). Examples (1) and (2) illustrate this difference in the use of maar:

(1) [Ik heb maar acht punten], **maar** [het is al 30 april].
A & B
*[I have only 8 coupons], **but** [it is April 30 already].*

(2a) [Het is nog geen 30 april], **maar** [ze zijn al wel op in de winkels].
A → B
*[It is not yet Apri l 30], **but** [ they are sold out already in the shops].*

(2b) **Hoewel** het nog geen 30 april is, zijn ze al wel op in de winkels.
***Although** it is not yet April 30, they are sold out already in the shops.*

Theoretically, an additive relation is a less complex than a causal relation is (cf. Evers-Vermeul & Sanders, 2009). Hence, in acquisition, the additive use of *maar* is expected to precede the causal use. Indeed, the results of the present study indicate that pupils with low writing proficiency have a tendency to use *maar* for a variety of relations, including additive (cf. Figure 33, Examples A and B). As writing ability increases the (incorrect) use of *maar*

decreases, and most occurrences of *maar* depict contrastive relations (cf. Figure 33, Examples C and D).

bij de winkel zijn geen repen meer met punten. dus ik wil ze opsturen met 2 zonder punten erbij **maar** het is nog geen 30 april **maar** die 2 repen die zijn wel heel hoor.
**maar** ik vind het wel jammer dat er geen repen meer zijn met geen punten. **maar** nou heb ik wel 8 punt **maar** dan 2 repen erbij zonder punten. **maar** ik kan er echt niets aan doen. **maar** ik wil de telefoon heel graag ontvangen.
ik heb maar 8 punten dus ik doe er maar 2 hele repen bij zonder punten **maar** ik vind het heel erg.

(A) Almost all relations indicated with *maar*

Ik heb maar acht punten, **maar** het is al 30 april. **Maar** ik eet altijd smikkelchocola. **Maar** ik ben drie weken op vakantie geweest en ik kwam gisteren terug dus, en ik las vandaag pas de actie. **Maar** ik wil zo graag die telefoon want mijn andere telefoon is kapot. Heb a.u.b. medelijden, ik wil die telefoon zo graag mag ik alstublieft die telefoon.
Als u mij een brief wil sturen stuur dan naar: Izaak Den Dekker. alvast bedankt en tot ziens

(B) A variety of relations indicated with *maar*

Ik heet: Ruben van giessen en spaar voor de gratis telefoon.
Ik heb al 8 zegels, **maar** de repen zijn op.
Het is nog geen 30 april, **maar** ze zijn op in alle winkels.
Daarom vraag ik u of ik die telefoon toch mag ontvangen.
Ik heb er echt alles aan gedaan om ze bij elkaar te krijgen.
Zou ik dus alstublieft die telefoon mogen ontvangen?

(C) Correct use of *maar* (negative additive and contrastive)

ik deed mee aan de actie van de gratis telefoon. Ik heb m'n best gedaan om de 10 punten bij elkaar te sparen **maar** dat is helaas niet gelukt. Omdat er in de supermarkt geen punten meer bij de Smikkelrepen zitten, en het is nog geen 30 april. Dus dat is niet zo eerlijk.
Ik kan er dus niks aan doen dat ik de 10 punten niet kan halen. Dus ik zou willen vragen of ik die gratis telefoon toch gratis opgestuurd zou mogen krijgen? Ik vind de Smikkel repen ook superlekker.

(D) Correct use of *maar* (contrastive)

*Figure 33.* Examples of the use of the connective *maar* (but).

Because the (negative additive) use of *maar* and the use of (positive) additive *en* (and) decrease with grade level, the present results support the hypothesis that the use of unelaborate connectives decreases with writing ability. However, to achieve a valid interpretation of the use of connectives as an indicator of writing ability, the exact relation between connectives and writing ability will have be investigated more thoroughly. First, since connectives differ in their conceptual and syntactical difficulty, the patterns of correct and incorrect use are likely to differ among connectives and across grades. Second, coherence does not depend solely on the use of explicit cohesive elements, such as connectives, but can be achieved more implicitly by the overlap between arguments and even without any (measurable) elements of cohesion. Hence, as is illustrated in Figure 33, connective density cannot simply be interpreted as a measure of writing ability. Instead, the correctness and relative difficulty of the connectives used should be taken into account when analysing text complexity as an indicator of writing ability.

Furthermore, text length was found to influence measures of coherence. Because less skilled writers produce shorter texts compared to highly able writers, a similar number

of cohesive elements, such as connectives, will induce a (relatively) high coherence density in shorter texts. In addition, longer texts written by proficient writers are likely to provide more information (e.g., several different reasons instead of one). This means that a (relatively) low number of arguments will be repeated within the text, thus lowering the amount of argument overlap, which is a measure of the coherence of content. These results showed the need to redefine the composite measures, such as by selecting only individual cases that show high agreement with writing ability or by assigning weights to individual cases, based on their correctness and relative complexity.

The relation between text complexity factors and aspects of writing ability was further investigated in an analysis of the outliers per measure. Based on these outliers, several factors that unjustly influence text complexity scores were identified, including task and stylistic elements. Text length is the most prominent factor influencing the degree of text complexity detected by T-Scan. With regard to density measures (e.g., prepositions) in T-Scan, outcomes are based on the number of occurrences per 1000 words. Hence, a very short text containing one or two prepositions receives a high complexity score for this measure, which may not be in line with the actual (low) complexity of the text. Furthermore, a longer text increases the chance of word repetition, which lowers scores on lexical diversity (type/token-ratio).

In an attempt to limit the influence of text length, an additional analysis was performed, in which several density measures were converted into averages per clause. However, the effect of this conversion on the outcomes of the text complexity measures (cf. Figure 22 through Figure 29), and the ordering of essays (cf. Table 18 through Table 20) was minimal. An alternative solution to the influence of text length would be to standardise text length, that is, to analyse a standard number of words per pupil. However, this would be difficult to realise in assessing novice writers because text length correlates strongly with writing ability. That is, because of differences in fluency, less able pupils will usually not be able to produce the same number of words as highly skilled writers can. An alternative would be to create a large corpus of words and/or sentences per grade and then to analyse the differences between random samples of these corpora.

Lastly, the interpretability of text complexity as an indicator of writing ability was evaluated by using the scores derived from T-Scan measurement as selection criteria to construct developmental scales of writing ability. Based on these text complexity scores, exemplars were selected to represent different grades and different ability levels, hence enabling an across-grade developmental comparison and a within-grade normative comparison of essays. The selected exemplars illustrated the developmental pattern of the selected text complexity features by demonstrating the increase in lexical sophistication and sentence complexity, and the difference in the use of (explicit) coherence elements. However, the process of selecting the exemplars confirmed the unwanted influence of several characteristics in the novice writers' texts (e.g., text length). That is, in constructing a scale, the validity of the selected exemplars should be confirmed by checking that their complexity

scores are indeed the results of linguistic features of complexity, instead of characteristics such as text length.

Given the aforementioned factors that (unjustly) influence text complexity, the future selection of exemplars cannot solely be based on empirical input because the complexity scores of an essay do not always reflect the actual complexity of the text (cf. Table 14 through Table 16). Instead, complexity scores can be used to select exemplar candidates, whereupon the essays that are wrongly classified should be rejected. Hence, a valid developmental scale can be produced to illustrate different levels of text complexity.

Showing the findings in this section of Chapter 4, Table 21 through Table 23 first provide a description of the selected measures of text complexity and their relation to writing ability, based on the agreement with grade level and score level. Second, the potential pitfalls that arise when using text complexity measures to evaluate the writing ability of novice writers are summarized, as well as their possible solutions.

Table 21. *Evaluation of Text Complexity Measures (Word Level)*

| **Measure** | **Description of measure** | **Relation to writing ability** |
|---|---|---|
| Word length | average number of letters per word | |
| Word complexity | average number of morphemes per word | lexical complexity: increases with writing ability |
| Word frequency | proportion of words that overlap with 50% most frequent words of word frequency list[1] | |
| Adverbials | average number of adverbials per clause | lexical richness: increases with writing ability |
| Type/token-ratio | average number of instances (types) per word (token) (av.) | |
| Prepositions | number of prepositions per 1000 words | |
| **Potential pitfalls and possible solutions** | | |
| task dependency | Certain words given within the task, can lower or raise average lexical complexity, without being indicative of writing ability. By specifying a task-specific list of words to be excluded from the analysis, this effect can be limited. | |
| spelling errors | Misspelled words cannot be mapped to word frequency lists and may not be recognized by a part-of-speech-tagger. Hence, they may be processed as low-frequent words and/or influence other measures on the word level. By using a pre-processing module in which spelling errors are detected and corrected, these effects are likely to be largely eliminated. | |
| text length | In a short text, the use of elements that indicate lexical complexity (e.g. long words) and/or lexical richness (e.g. prepositions) has a relatively strong influence on density measures of text complexity. In addition, the chance of using several tokens of the same type will increase with text length, thus lowering type/token-ratio, a measure of lexical diversity. Computing averages per clauses (instead of density) does not seem to affect the measures greatly (cf. Figure 22 and Table 18). By further investigating the specific effect of text length on word level measures, and/or standardising the number of words on which the measures are based, these effects can possibly be corrected for. | |

[1] Staphorsius, 1994

Table 22. *Evaluation of Text Complexity Measures (Sentence Level)*

| Measure | Description of measure | Aspect of writing ability |
|---|---|---|
| Subordinate clauses | number of subordinate clauses (per 1000 words) | sentence complexity: increases with writing ability |
| Dependency length | average distance between sentence components* | |
| Clauses | number of clauses per 1000 words | |
| Sentence length | average number of words per sentence | |
| **Potential pitfalls & possible solutions** | | |
| lack of punctuation | In texts were (almost) no periods are used to indicate sentence endings, sentence length will be extremely high, thus influencing other measures based on sentence length. By using a pre-processing module in which either only those essays lacking punctuation are detected and corrected, or punctuation is standardised across all essays, this effect is likely to be largely eliminated. | |
| spelling errors | Misspelled words may not be recognised by the grammatical parser, hence possibly causing errors in the grammatical parsing of sentences, and influencing measures based on grammatical parsing. By using a pre-processing module in which spelling errors are detected and corrected, these effects are likely to be largely eliminated. | |
| text length | In a short text, the use of complex sentence structures has a relatively strong influence on sentence complexity. By further investigating the specific effect of text length on sentence level measures, and/or standardising the number of words on which the measures are based, these effects can possibly be accounted for. | |

* subject-verb; direct object-verb; object-verb; verb-preposition; determiner-noun; preposition-noun; finite verb-main verb ; subordinate conjunction-finite verb subordinate clause; coordinating conjunction-conjunct head; subordinate conjunction-main verb; noun-subordinate clause

Table 23. *Evaluation of Text Complexity Measures (Text Level)*

| Measure | Description of measure | Aspect of writing ability |
|---|---|---|
| Argument overlap | overlap between (lemmatized) arguments* in consecutive sentences per 1000 words | (explicit) coherence: decreases with writing ability |
| Referential pronouns | number of 3rd person personal and possessive pronouns and demonstrative pronouns per 1000 words | |
| Connectives | number of different categories** of connectives per 1000 words | |
| **Potential pitfalls and possible solutions** | | |
| homonyms | Homonyms of connectives are sometimes classified as connective (and vice versa). Exclusion of these words, and/or improvement of part-of-speech and grammatical parsing can reduce the numbers of wrongly classified words. | |
| correctness | Not all cohesive devices are used in a grammatically or conceptually correct manner. Hence, their relation to writing ability depends on their correctness within the context. By only incorporating measures that are relevant given the specific task, and by using a pre-processing module in which grammatical errors are detected and corrected, these effects can possibly be accounted for. | |
| task | The use of (explicit) coherence relations partly depends on the text goal that is elicited by the task given. Hence, the absence or presence of a specific relation can only be valued when taking into account the text goal. By only incorporating the relevant measures per task, this effect can possibly be accounted for. | |
| text length | In a short text, the use of cohesive elements has a relatively strong influence on the coherence measures. Computing averages per clauses (instead of density) does not seem to limit the effect of text length (cf. Tables 24, 26 and 28; Figures 19 and 20). By further investigating the specific effect of text length on coherence measures, and/or standardising the number of words on which the measures are based, these effects can possibly be corrected for. | |

* arguments: pronouns (excl. demonstratives); proper nouns; nouns; main verbs
** temporal; enumerative; contrastive; comparative; causal

## 4.4 Discussion and conclusion

This study explored the applicability of automated essay evaluation (AEE) within a large-scale assessment in primary education. For this purpose, the validity of text complexity measures as indicators of writing ability was assessed. To be eligible as an indicator of writing ability, text complexity measures must be both meaningful and interpretable as an assessment of text quality, as well as suitable as indicators of writing development. Because no previous studies have investigated the relation between automatically generated measures of text complexity and writing ability in Dutch, this study aimed at taking the initial steps towards building an AEE instrument.

First, the specific text complexity measures that are indicative of writing ability were identified by evaluating their agreement with both grade level and human essay scores. This result was used in a selection of 13 complexity measures that represent lexical sophistication, sentence complexity and coherence. (Section 4.2.). To gain insight into the use of text complexity measures in a descriptive evaluation of writing performance, a qualitative analysis of the selected measures was performed (Section 4.3). For measures comprised of a fixed selection of words (e.g., connectives), an analysis of the individual instances was performed. In addition, to identify characteristics of novice writing that unjustly influence complexity scores—thus impairing validity—an analysis of outliers was performed. Finally, based on AEE outcomes, a developmental scale of writing ability was created and illustrated using exemplars of written products that represented different age groups and ability levels.

Pre-processing module

The present study showed that novice writers produce texts in which spelling and punctuation is flawed to a relatively great extent. These findings are supported by outcomes of the national assessment on writing quality in The Netherlands (Van Til et al., 2014). In the third grade of primary education (Dutch grade 5), pupils misspell 19 per cent of verbs and 8 per cent of other words, while 30 per cent of pupils produce texts without any punctuation. In the final grade of primary education (6th grade, Dutch 8th grade), the writing quality is improved: 13 per cent of pupils do not use punctuation elements, 8 per cent of verbs are misspelled, and 3 per cent of other words contain spelling errors (Van Til et al., 2014).

The analysis of essays written by novice writers in this study demonstrated that errors in spelling and lack of punctuation negatively influence the validity of an automated evaluation of writing. First, misspellings caused errors in the identification of words, which consequently received an incorrect part-of-speech tag and/or word frequency score. Second, because the grammatical parser relies on sentence stops in identifying a sentence, the lack of punctuation causes the parser to process large pieces of unpunctuated text as one, highly complex, sentence.

For the use of AEE in populations of novice writers, a "pre-processing" module (i.e., Pre-Scan) should therefore be developed. Incorporated in an automated evaluation of

writing quality, this module would automatically *detect* and *correct* writing products in which spelling and punctuation is flawed, in order to be reliable and valid evaluation of writing ability. Firstly, Pre-Scan would detect and automatically correct spelling errors, including spacing errors and misspelled verbs. Furthermore, the Pre-Scan module should be able to detect writing products that lack punctuation and should consequently insert a minimal number of sentence stops. Previous research on the automated insertion of punctuation elements in an English literary text using a data-driven approach indicated the feasibility of a module of this kind (Krahmer & Van den Bosch, 2006). However, the fact that an abundant use of sentence stops influences sentence length, and hence complexity measures, (cf. Figure 34, Example A), speaks for the standardisation of punctuation across essays by replacing all punctuation with a rationalised minimal quantity of punctuation elements (cf. Figure 34, Example B).

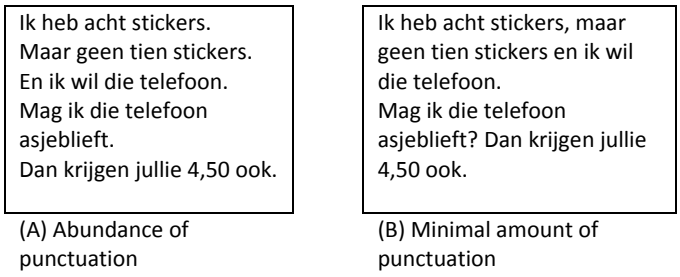



*Figure 34.* Differences in punctuation style.

Although the need to edit the writing products of novice writers prior to analysing their writing quality may be seen as an undesirable precondition, a pre-processing model also offers benefits. First, removal of unwanted elements, such as letter headings, could be incorporated into Pre-Scan, as well as a check on the faithfulness and authenticity of the essays. Furthermore, information gathered in the Pre-Scan module could be used to evaluate spelling and punctuation as specific aspects of writing. This way, a pre-processing module would ensure that the necessary conditions for automated evaluation were satisfied, while providing an additional source of information on writing quality at the same time. Finally, by using a pre-processing module, the inability of human raters to ignore spelling errors and evaluate other aspects of writing independently would be overcome.

The validity and interpretability of text complexity measures as indicators of writing ability

The validity of automated essay evaluation depends on the inferences that are intended to be drawn from the outcome (Clauser et al., 2002). The Dutch national assessment aims to describe group differences in writing performance. AEE is considered a reliable method in comparing groups within a low-stakes assessment (Deane & Quinlan, 2010; Shermis & Hamner, 2013). To support validity further, AEE models should be tailored for a specific use,

that is, specific models should be built for specific writing prompts and/or populations (Huot, 1996; Yang et al., 2002; Deane & Quinlan, 2010; Ramineni, 2013).

As part of the validation process, the correspondence between AEE and the intended construct has to be demonstrated (Clauser et al., 2002). In the present study, this correspondence was evaluated by means of agreement between text complexity scores and both grade level and human essay scores. This agreement was evaluated in a body of essays written in response to a single task. Hence, to generalise these findings across tasks, topics, and text genres, the results of different writing prompts will also have to be investigated. Based on these results, conclusions can be drawn regarding the generalisability of the AEE-results, which provide further insight into the validity of this method in the evaluation of writing ability.

As Crossley et al. (2011) stated, linguistic features are the most salient features that can be quantitatively measured. The use of linguistic features as indicators of writing ability is based on the notion that text quality can be evaluated by examining the linguistic structures of a text (cf. Hayes & Flower, 1980; McCutchen & Perfetti, 1982; Perfetti & McCutchen, 1987, Sanders & Van Wijk, 1996; Van Wijk & Sanders, 1999). However, the exact relation between linguistic features and text quality has to be determined for each language and population to which AEE is applied. The present study showed that in Dutch, several features of linguistic complexity are related to writing ability. Table 24 provides a selection of the features of linguistic complexity that were found to indicate writing ability.

Table 24. S*elected Complexity Measures for an Automated Evaluation of Writing Ability*

| | **Measure** | **Description** | **Suggested relation to writing ability** |
|---|---|---|---|
| WORD | | | |
| **Lexical complexity** | word length | number of letters per word (average) | + increases with ability |
| | word complexity | number of morphemes per word (average) | + increases with ability |
| | word frequency (50%) | proportion of words that overlap with 50% most frequent words of word frequency list | - decreases with ability |
| **Lexical richness** | adverbials | number of adverbials per clause (average) | + increases with ability |
| | type/token-ratio | number of different instances (tokens) per total number of lemmas (types) (average) | - decreases with ability |
| | prepositions | number of prepositions (per 1000 words) | + increases with ability |
| SENTENCE | | | |
| **Sentence complexity** | subordinate clauses | number of subordinate clauses (per 1000 words) | + increases with ability |
| | dependency length | distance between sentence components (average) | + increases with ability |
| | clauses | number of clauses (per 1000 words) | - decreases with ability |
| | sentence length | number of words per sentence (average) | + increases with ability |
| TEXT | | | |
| **Coherence** | argument overlap (lemmabuffer) | overlap between (lemmatized) arguments in preceding 10 words (per 1000 words) | - decreases with ability |
| | referential pronouns | number of 3rd person personal/possessive pronouns and demonstrative pronouns (per 1000 words) | - decreases with ability |
| | connectives | number connectives (of different categories) (per 1000 words) | - decreases with ability |

In addition, this study revealed the influence of several text characteristics on the validity of AEE measures. First, spelling errors and lack of punctuation cause words to be misclassified and pieces of unpunctuated text to be unjustly evaluated as complex sentences, thus impairing validity. Validity is further impaired by the limited text length of essays written by novice writers because measures that are based on the density of features (i.e., their occurrence per 1000 words) are strongly influenced by text length. For example, a short text in which one or two prepositions are used will receive a relatively high value for this measure, compared to a longer text. Hence, text length increases text complexity scores, while text length itself is negatively correlated to writing ability. To determine a method to overcome this unwanted effect, an additional analysis was performed in which average numbers of prepositions, referential pronouns and connectives were computed per clause instead of per 1000 words (density). The results of this analysis indicated that conversion from density measures to measures per clause did not limit the influence of text length. However, the conversion might well serve to limit the effect of flaws in punctuation because the identification of clauses does not rely on the presence of sentence stops.

In addition, the type/token ratio is a well-known measure for diversity in vocabulary: a high type/token ratio implies that the several tokens (words) used in an utterance belong to different types, that is, many different words are used. In longer texts, however, the number of words that are repeated naturally increases. A similar effect was described by Lee, Gentle, and Kantor (2009). Several methods to reduce the effect of text length have been proposed in the literature, including analyzing a random set of tokens or sequences of words. Koizumi (2012) studied the use of these measures to evaluate short texts and concluded that the effect of text length can be reduced when analyzing texts with a minimum of 100 tokens.

In addition to the effects caused by text characteristics, other influences are likely to affect the quality of AEE measures. In T-Scan, several corpora of written language are used to provide word frequency and other word features. However, the content of these corpora is based on adult language and will therefore not fully represent the language use of novice writers. In addition, several measures in T-Scan compile a collection of words that are believed to represent one textual feature (e.g., referential pronouns and connectives). The analysis of these measures imply that some instances within these measures might indicate high ability, whereas other instances within the same measure indicate low ability, resulting in a measure that is both positively and negatively related to writing ability. For example, results of the qualitative analysis (Section 4.3) suggested that the negative additive use of *maar* decreased with grade level, whereas causal use indicated high writing ability (cf. Figure 33).

Figure 35 shows essays of low and high quality, illustrating the various uses of connectives. These examples demonstrate the need to examine further the use of different types of connectives by novice writers, the effects of text length, and the relation of connective use to coherence as an aspect of text quality. Because coherence is believed to be

a critical, yet ill-defined characteristic of text (Burstein, Tereult, Chodrow, Blanchard, & Andreyev, 2013), the results of this further analysis would be a valuable contribution to the field of language assessment.

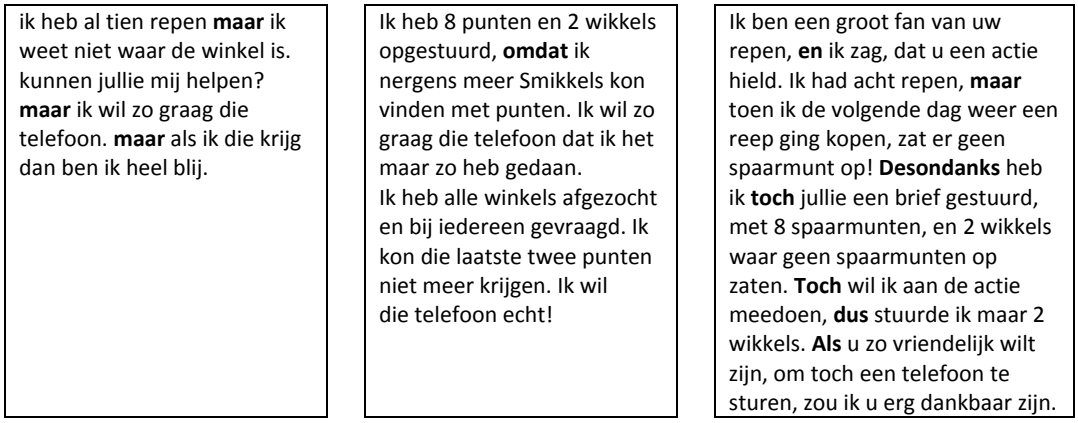



*Figure 35.* Examples of the use of connectives.

To evaluate the interpretability of text complexity measures as indicators of writing ability, exemplar essays were selected based on the selected measures from T-Scan. Developmental scales for linguistic features at the level of word, sentence and textual level were composed by selecting essays based on their complexity scores (cf. Figure 22 through Figure 24). Although selection of the essays required attention to the possibility of factors unjustly influencing complexity scores, the scales demonstrated that the complexity measures in T-Scan could indeed be interpreted as indicators of writing ability. The exemplars illustrated development in writing ability by demonstrating an increased level of lexical sophistication and sentence complexity and a decreased density of cohesive elements.

The applicability of AEE within a large-scale assessment in primary education

In general, AEE can be used in several ways within a (large-scale) assessment, including as a substitute for human scores, a check on human scores, or an independent second measure. Using AEE scores as an efficient method to substitute human scores is considered a weak argument for adopting AEE, because a basic measure, such as essay length, might be sufficient to predict human scores reliably (Lee et al., 2009; Deane & Quinlan, 2010; Williamson, 2013). Instead, AEE could prove valuable when used as a *complement* to human scores because humans and machines are suited to evaluating different aspects of writing. In a combined human–AEE scoring method, human raters should therefore focus on evaluating high-order skills, whereas AEE should be used as a specific and consistent measure of certain

linguistic features (Williamson, 2013, Attali, 2013). Hence, full representation of constructs is ensured and meaningful linguistic surface features are independently evaluated.

Used as a complement to human ratings, AEE offers a measure on the linguistic level, which is the estimated zone of proximal development for novice writers. Most pupils in grade 3 will have mastered writing skills at the preceding neural level (e.g., production of letters), while pupils up to grade 6 will still be constricted at the linguistic level, preventing a developmental focus on the cognitive level (e.g., planning) (Abbott & Berninger, 1993). Furthermore, AEE enables a fine-grained analysis of large corpora and provides judgments that are comparable across time, raters and tasks. By offering a consistent metric, AEE is considered especially suitable as a methodology in trend studies, in which writing performance across cohorts is compared, such as in national assessments (Deane & Quinlan, 2010; Williamson, 2013).

In this study, the use of text complexity measures as indicators of writing ability was evaluated. The results presented in Table 7 and Table 8 show that AEE measures are capable of characterizing group differences (cf. Deane & Quinlan, 2010), as well as offering a yardstick and context for interpreting performance levels across grade levels and trends in performance over the years. In addition to its applicability within a (large-scale assessment), the multi-trait scoring of writing ability within AEE can serve other evaluative goals, such as providing insight into the assessment criteria used by teachers and offering feedback to learners (cf. Swartz et al., 1999).

Based on the results described in this chapter, the text complexity measures provided by T-Scan seem potentially useful for an evaluation of writing ability. However, several factors were found to influence unjustly indicators of text complexity, especially in the assessment of the writing products of novice writers, which are typically flawed and short. These factors can possibly be accounted for by analysing samples of individual words and/or sentences per group instead of complete essays. This way, text length is eliminated as an unwanted influence on the text complexity score. When unwanted influences are acconted for, AEE seemes a useful application within a national assessment of writing ability, the goal of which is to monitor (development in) writing ability by identifying and explaining differences in performances between groups of pupils.

Future research

A major issue in writing assessment is the fact that no “golden standard” of what constitutes good writing quality is available (Shermis et al., 2013). Future research should thus focus on acquiring a better understanding of the development of writing performance. AEE can support this investigation by providing a systematic method to investigate writing products that represent different levels of writing performance (Attali & Burstein, 2006). By analysing a corpus of essays written by a particular population, group differences can be characterized in terms of specific aspects of writing. The results of AEE could then help define the characteristics of high and low writing quality at different developmental stages within this

population, providing exemplars that clarify the concept of writing quality (Deane & Quinlan, 2010) and illustrating particular features of text quality. To improve the understanding of writing development and writing quality, the techniques and procedures that underlie AEE methods need to be disclosed to writing experts, language teachers and the public (Lee et al., 2009; Cushing Weigle, 2013; Ramineni & Williamson, 2013).

In addition to providing information about the writing product, AEE offers an opportunity to gather information about the writing process. By incorporating key-stroke logging into AEE-systems, information about the process of producing a text is revealed (Deane & Quinlan, 2010; Almond, Quinlan & Attali, 2011; Deane, 2013). Key-stroke logging offers information about text production, such as production speed and fluency, the order in which elements are produced, and the amount and type of revisions that are executed while writing (cf. Leijten & Van Waes, 2013). Hence, AEE enables the evaluation of certain aspects of the process, which are believed to influence writing quality and therefore reflect writing ability, but which may not be revealed by solely evaluating the final written product.

Hence, in addition to the applicability of AEE in large-scale assessments, the automated evaluation of writing offers the opportunity to gather valuable information for educational purposes, thus providing support to language learners as well as language teachers. Monitoring the writing process in AEE could disclose information about the process of text production. Furthermore, qualitative research into the specific textual features that indicate writing quality could gain insight into the components that comprise writing ability.

**Conclusion**

The present study explored the possibility to use automated essay evaluation (AEE) in the Dutch national assessment in primary education. Because no AEE application is readily available for the Dutch language, T-Scan, a program for the automated evaluation of text complexity in Dutch, was used. The results revealed that flawed writing products influence the validity of numerous text complexity measures, which highlighted the need to develop a pre-processing module (Pre-Scan), in which errors in spelling and punctuation are detected and subsequently corrected.

Furthermore, the analysis of agreement between 37 complexity measures from T-Scan and both grade level and human essay scores, resulted in the selection of 13 measures of text complexity that indicate writing ability. First, the evaluation of their validity and interpretability as measures of writing ability highlighted the need to adjust composite measures (e.g., connectives) so that they are specifically suited for evaluating the writing ability of novice writers. Second, the need to limit the influence of factors that unjustly influence complexity scores (e.g., text length) became evident. Finally, the constructed developmental scales demonstrated that exemplars selected by means of complexity measures of the word, sentence and text levels can be employed to depict the development of novice writers.

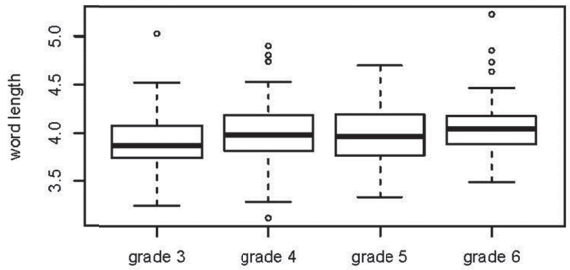
In conclusion, the results of this study indicated the need for further research to substantiate the usability of AEE within primary education. First, the relation between text features and writing ability has to be explored further. Second, the effects of characteristic features of novice writers' texts on the validity of the evaluation of these features should be specified. Nonetheless, in its current state, AEE is capable of offering a consistent metric and a fine-grained analysis of quality traits related to linguistic surface characteristics of essays. Given these properties, AEE can be applied to characterize group differences across particular traits in large-scale tests. It therefore can support writing education by offering an objective and informative method to complement the assessment of writing ability.

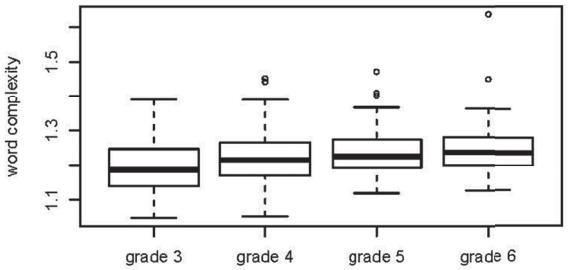
# Appendix A – Feature Variables in T-Scan (March 2013)

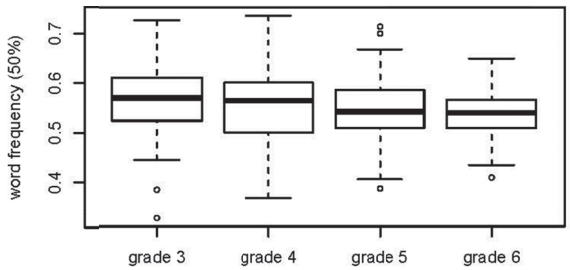
| | | | |
|---|---|---|---|
| **WORD COMPLEXITY** | | | |
| N=24 | lpw | sdpw | lemma_freq_log |
| | wpl | sdens | lemma_freq_log_zn |
| | lpwzn | freq50 | freq1000 |
| | wplzn | freq65 | freq2000 |
| | mpw | freq77 | freq3000 |
| | wpm | freq80 | freq5000 |
| | mpwzn | word_freq_log | freq10000 |
| | wpmzn | word_freq_log_zn | freq20000 |
| **SENTENCE COMPLEXITY** | | | |
| N=30 | wpz* | nom | deplen_verbpp* |
| | pzw* | lv_d | deplen_noundet* |
| | wnp | lv_g | deplen_prepobj* |
| | subord_clause | prop_negs | deplen_verbvc* |
| | rel_clause | morph_negs | deplen_compbody* |
| | clauses_d | total_negs | deplen_crdcnj* |
| | clauses_g* | multiple_negs* | deplen_verbcp* |
| | dlevel* | deplen_subverb* | deplen_noun_vc |
| | dlevel_gt4_prop* | deplen_dirobverb* | deplen |
| | dlevel_gt4_r* | deplen_indirobverb* | deplen_max* |
| INFORMATION DENSITY | | | |
| N=12 | word_ttr | content_words_g | np_mods_d* |
| | lemma_ttr | rar_index | np_mods_g* |
| | content_words_r | vc_mods_d* | np_dens |
| | content_words_d | vc_mods_g* | conjuncts* |
| COHERENCE | | | |
| N=17 | temporals | argument_overlap_d* | lemmabuffer_argument_overlap_d |
| | reeks | argument_overlap_g* | lemmabuffer_argument_overlap_g |
| | contrast | lem_argument_overlap_d* | indef_nps_p |
| | comparatief | lem_argument_overlap_g* | indef_nps_r |
| | causal | wordbuffer_argument_overlap_d | indef_nps_g |
| | referential_prons | wordbuffer_argument_overlap_g | |
| CONCRETENESS | | | |
| N=12 | noun_conc_strict_p | noun_conc_broad_r | adj_conc_strict_d |
| | noun_conc_strict_r | noun_conc_broad_d | adj_conc_broad_p |
| | noun_conc_strict_d | adj_conc_strict_p | adj_conc_broad_r |
| | noun_conc_broad_p | adj_conc_strict_r | adj_conc_broad_d |
| PERSONALITY | | | |
| N=23 | pers_ref_d | action_verbs_p | emo_adjs_p |
| | pers_pron_1 | action_verbs_d | emo_adjs_d |
| | pers_pron_2 | state_verbs_p | imperatives_p |
| | pers_pron3 | state_verbs_d | imperatives_d |
| | pers_pron | process_verbs_p | questions_p |
| | names_p | process_verbs_d | questions_d |
| | names_r | human_nouns_p | polarity |
| | names_d | human_nouns_d | |
| PARTS OF SPEECH | | | |
| N=10 | adj | vz | noun |
| | vg | bijw | verb |
| | vnw | tw | interjections |
| | lid | | |
| MISCELLANEOUS | | | |
| N=19 | present_verbs_r | copula_g | infin_d |
| | present_verbs_d | archaics | infin_g |
| | modals_d | vol_deelw_d | surprisal |
| | modals_g | vol_deelw_g | wopr_logprob |
| | time_verbs_d | onvol_deelw_d | wopr_entropy |
| | time_verbs_g | onvol_deelw_g | wopr_perplexity |
| | copula_d | | |

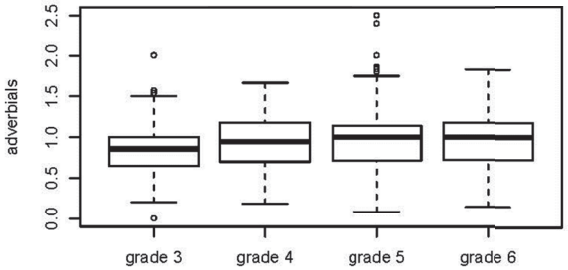
*value is influenced by sentence length

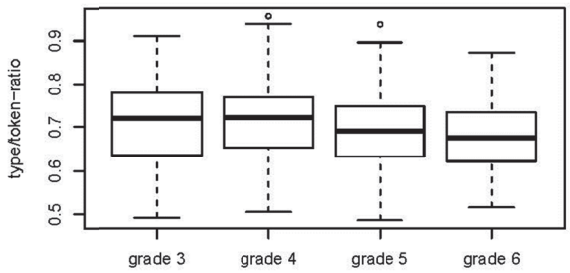
## Appendix B – Feature Values per Grade Level

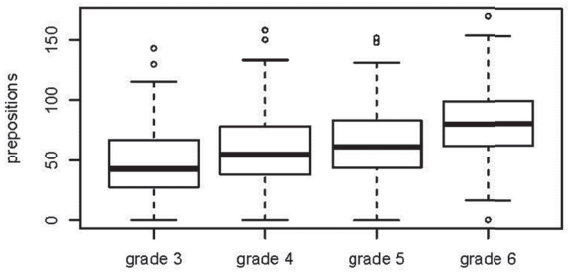


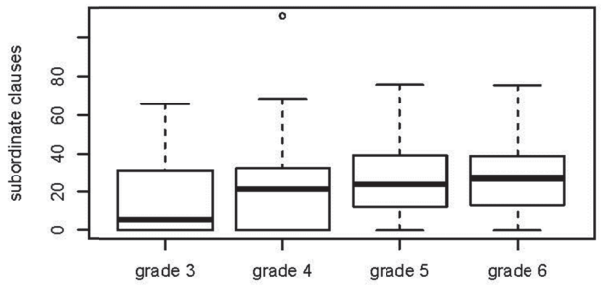



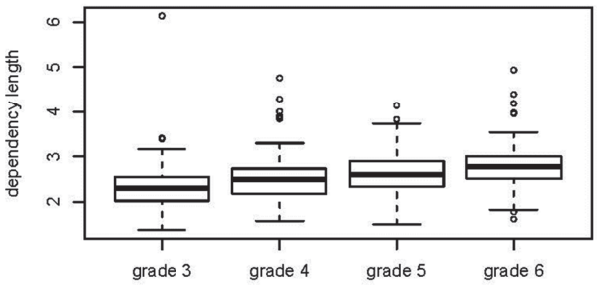



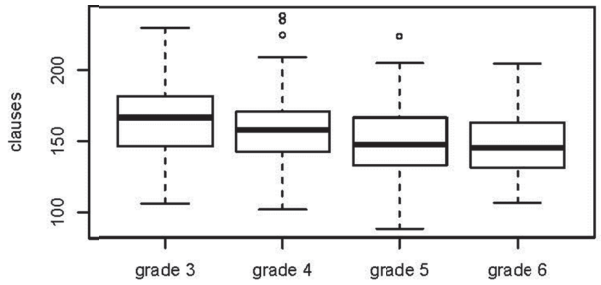



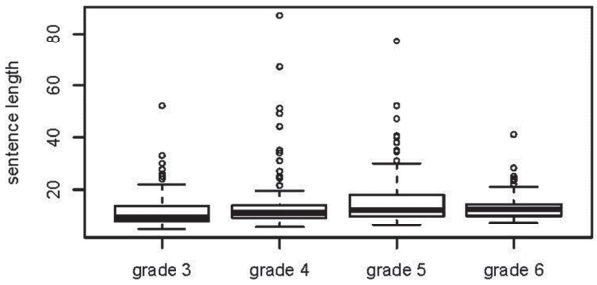



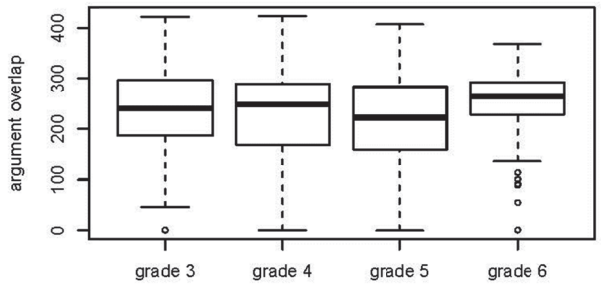



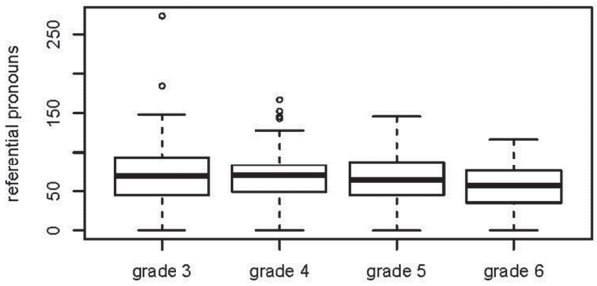

subordinate clauses
0
20
40
60
80
grade 3
grade 4
grade 5
grade 6


dependency length
2
3
4
5
6
grade 3
grade 4
grade 5
grade 6

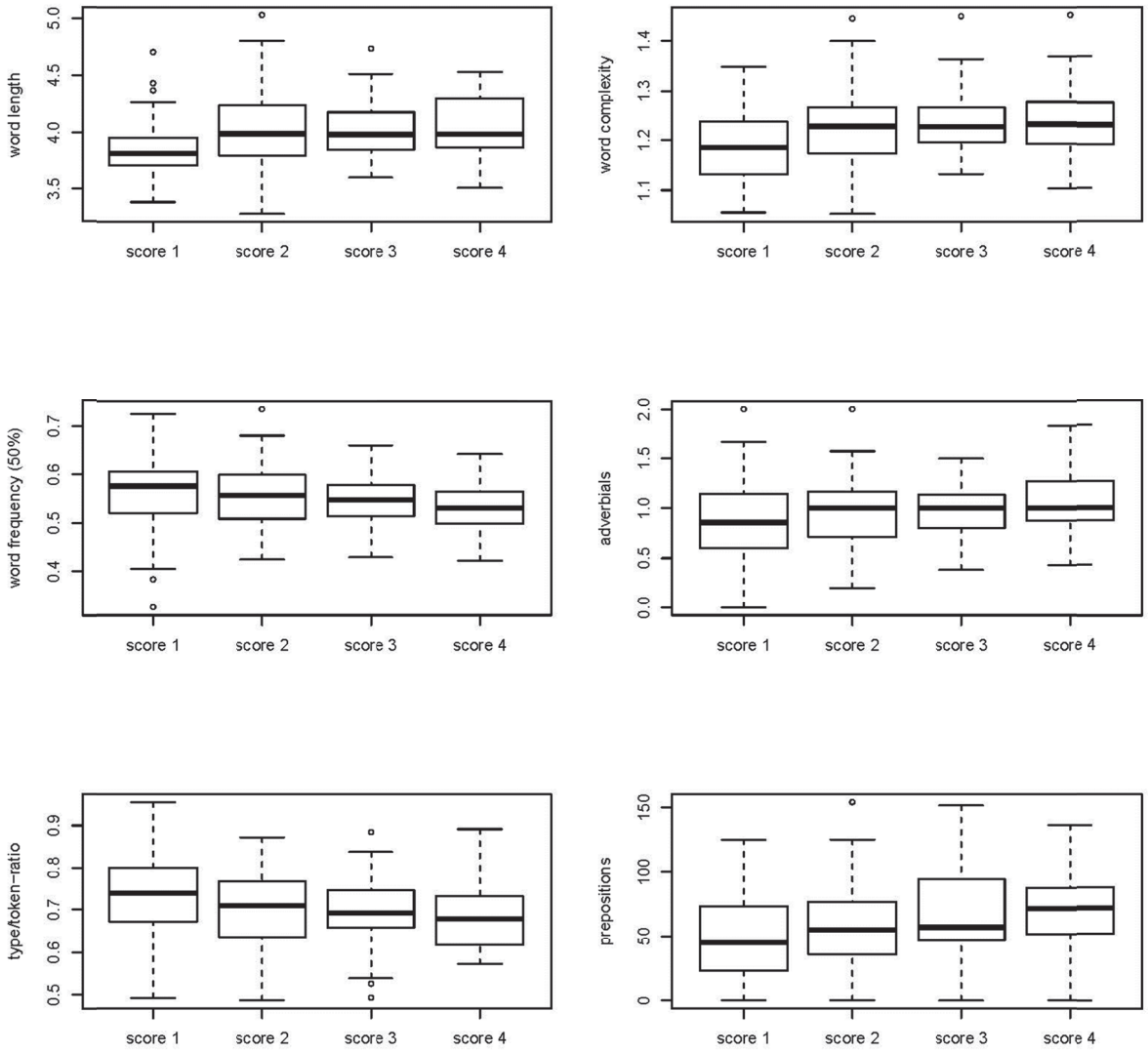

clauses
100
150
200
grade 3
grade 4
grade 5
grade 6

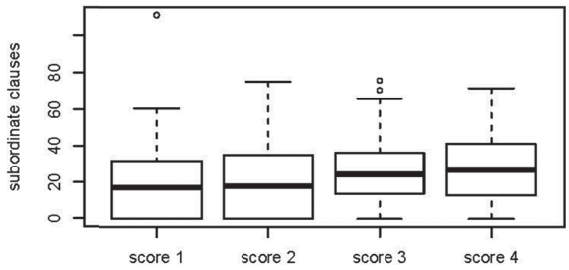

sentence length
20
40
60
80
grade 3
grade 4
grade 5
grade 6

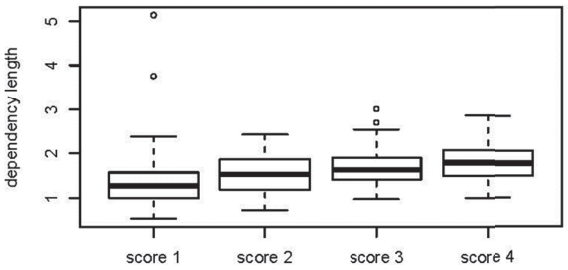

argument overlap
0
100
200
300
400
grade 3
grade 4
grade 5
grade 6

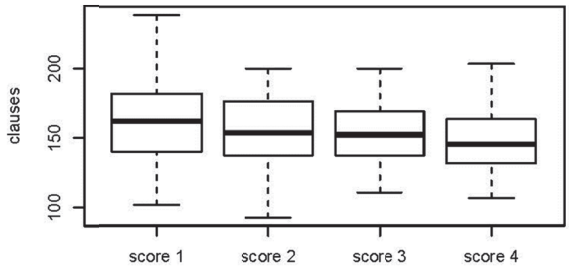

referential pronouns
0
50
150
250
grade 3
grade 4
grade 5
grade 6

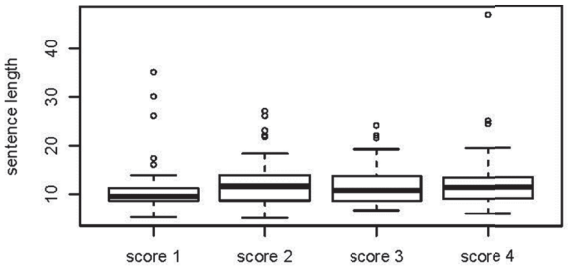
## Appendix C – Feature Values per Score Level

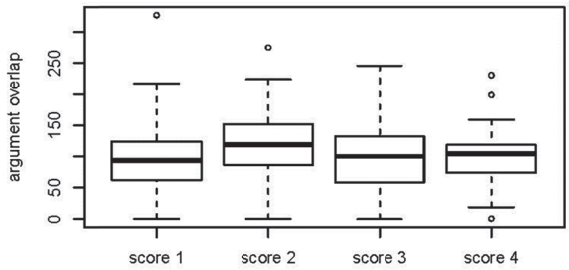


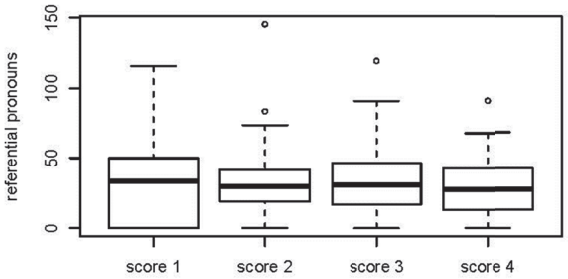

subordinate clauses
0
20
40
60
80
score 1
score 2
score 3
score 4


dependency length
1
2
3
4
5
score 1
score 2
score 3
score 4


clauses
100
150
200
score 1
score 2
score 3
score 4


sentence length
10
20
30
40
score 1
score 2
score 3
score 4


argument overlap
0
50
150
250
score 1
score 2
score 3
score 4


referential pronouns
0
50
100
150
score 1
score 2
score 3
score 4

## Appendix D – Prepositions, Connectives and Referential Pronouns: Number of Occurrences per Type

| Preposition | N | Preposition | N | Referential pronoun | | N |
|---|---|---|---|---|---|---|
| achter | 1 | toe | 6 | hen | them | 1 |
| beneden | 1 | door | 8 | ie | he | 1 |
| heen | 1 | over | 9 | zij | she/they | 1 |
| na | 1 | uit | 13 | haar | her | 2 |
| ondanks | 1 | tot | 21 | 'm | him/it | 3 |
| onder | 1 | mee | 39 | zijn | his | 3 |
| sinds | 1 | om | 60 | zo'n | such a | 10 |
| tegenover | 1 | zonder | 68 | hij | he | 19 |
| vanaf | 1 | naar | 73 | hier | here | 20 |
| via | 1 | te | 100 | daar | there | 40 |
| per | 2 | van | 101 | hem | him | 40 |
| rond | 2 | voor | 120 | dit | this | 45 |
| tussen | 2 | met | 125 | deze | this/these | 59 |
| af | 3 | aan | 127 | ze | they/she | 106 |
| langs | 3 | bij | 145 | dat | that/which | 116 |
| tegen | 3 | in | 202 | die | that/which/who | 279 |
| als | 4 | op | 240 | | | |
| binnen | 4 | **total** | **1490** | | **total** | **745** |

| Connective | N | Connective | N | Connective | N |
|---|---|---|---|---|---|
| behalve | 1 | opeens | 2 | waarom | 14 |
| binnenkort | 1 | aangezien | 3 | dan | 18 |
| daarmee | 1 | daardoor | 3 | terwijl | 21 |
| desondanks | 1 | doordat | 3 | tot | 21 |
| erom | 1 | eerst | 3 | Alvast | 25 |
| hiermee | 1 | wanneer | 3 | of | 47 |
| hiervoor | 1 | zelfs | 3 | toen | 48 |
| hoelang | 1 | daarvoor | 4 | om | 57 |
| na | 1 | in | 4 | al | 64 |
| nadat | 1 | meteen | 4 | maar | 74 |
| ondanks | 1 | namelijk | 4 | nu | 86 |
| sinds | 1 | zodat | 4 | omdat | 87 |
| straks | 1 | zo | 5 | daarom | 103 |
| vandaar | 1 | zoals | 5 | ook | 106 |
| vanochtend | 1 | hoewel | 6 | als | 108 |
| voordat | 1 | steeds | 6 | want | 138 |
| waarmee | 1 | vandaag | 7 | dus | 199 |
| gisteren | 2 | net | 10 | en | 567 |
| morgen | 2 | anders | 11 | | |
| ondertussen | 2 | pas | 12 | **total** | **1908** |

# 5 Discussion and Conclusion



## 5.1 Text quality and text complexity as indicators of writing ability

Producing a written text is a multifaceted process in which a writer engages in several complex cognitive activities, such as planning what message to communicate, mapping language onto thoughts, and monitoring the quality and appropriateness of the text produced thus far (Deane et al., 2008). Within a writing assessment, a candidate typically demonstrates his or her ability by executing one or more writing tasks. The quality of the resulting writing product(s) is subsequently evaluated by one or more raters, resulting in a judgement on the candidate's writing performance.

However, it is difficult to generalise such a judgement across the range of all possible tasks and raters to the construct of writing ability. A writing task requires a writer to produce a text within a fixed communicative setting, serving a specific rhetorical goal and aiming at a certain intended audience. Because of these (intertwined) facets of writing ability, the evaluation of a written product is a notoriously complex task for a rater to perform conscientiously and consistently. Typically, raters differ in their ideas as to which textual elements demonstrate writing ability, and are prone to inconsistent behaviour when applying evaluation criteria—giving rise to impaired reliability and validity of writing scores. These rater characteristics highlight the importance of providing raters with an unambiguous rating instruction in order to facilitate objective scoring.

In the present study, three different approaches to assess writing ability in primary education were evaluated. By doing so, this study aimed at providing recommendations to improve the validity and reliability of the assessment of novice writers in general, and the assessment of writing ability within the Dutch national assessment in primary education in particular. In the introductory Chapter 1, elements of the writing product were related to specific components of the writing process (Figure 1). The subsequent chapters discussed different methods to assess these components, focusing on both text quality (Chapters 2 and 3) and text complexity (Chapter 4). Together, these chapters aimed to answer the central question of this dissertation: *How can the writing ability of novice writers be assessed reliably and validly?*

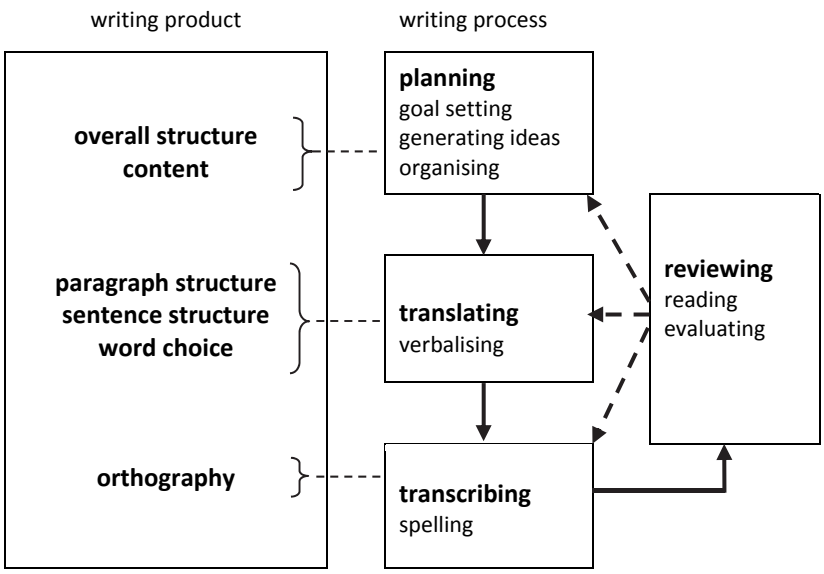



*Figure 1.* Elements of the writing product mapped to components of the writing process. (based on Flower & Hayes, 1981; Hayes, 2012; Van der Pool, 1995)

The three methods that were successively addressed in this study are the addition of anchor essays to an analytical assessment (anchored analytical assessment [AAA]) in Chapter 2, the assessment of revision ability (revision test [RT]) in Chapter 3, and the use of automated evaluation techniques (automated essay evaluation [AEE]) in Chapter 4. The results presented in these chapters imply that the methods applied within this study cover different stages in the practice of producing a text (i.e., the writing *process*), by evaluating different parts of the results of this process (i.e., the writing *product*).

Moreover, the present study combined two approaches to evaluate writing ability by addressing both the *quality* and the *complexity* of written texts. This way, information was gathered on the relation between specific features of a written product on the one hand, and the ability of its writer on the other. Although text complexity can be considered an aspect of text quality in some contexts, this relation is valid only if complexity is demanded by the specific communicative setting and rhetorical goal of the text produced. In the present study however, text complexity is deployed to provide a measure of writing ability that is independent of the context that is required by the writing task.

In the present section, the results of the methods applied in Chapter 2 through Chapter 4 are further evaluated by discussing the use of both text quality (Section 5.1.1) and text complexity (Section 5.1.2) as indicators of writing ability.

### 5.1.1 Evaluating text quality

The assessment of a candidate's writing ability is usually operationalised by evaluating the quality of one or more texts written by the candidate. On the basis of the quality of this performance, conclusions are drawn on the writer's ability. However, the precise relation between text quality and writing ability is tentative. On the one hand, quality criteria are largely dependent on the communicative setting and rhetorical goal of the writing task; a feature such as originality is usually highly valued when writing a fictional story, while the same feature might in fact lower the quality of a formal letter. Hence, text quality cannot be determined without taking into consideration the intended purpose of the written product. On the other hand, the rating of essay quality is known to be influenced by factors other than the candidate's ability, such as the specific task, rater, and rating procedure that are assigned. In other words, the validity of text quality criteria depends on the specific writing context, while reliability and generalisability are impaired by unwanted sources of influence on the quality ratings.

To improve the validity and reliability of a large-scale writing assessment in primary education, two assessment procedures were altered and evaluated in this study, considering both a reader's and a writer's perspective on text quality. When evaluating writing, a rater has to value the quality of a text from the viewpoint of the intended reader, taking into account the specific communicative setting and rhetorical goal imposed by the writing task. When producing a text, a writer engages in a similar evaluative task by reviewing the text produced and—if necessary—revising textual elements in order to improve text quality.

Chapter 2 evaluated the use of an anchored analytical assessment (AAA) procedure. In this newly developed rating procedure, a rating scale with anchor essays was combined with analytical questions. Three different aspects of text quality (i.e., content, structure, and correctness) were evaluated by means of both a list of analytical questions and a comparison to exemplar essays representing different ability levels. This way, detailed information on specific text features was collected; at the same time, attention to the wholeness of the written product was guaranteed. Results of the AAA procedure indicated that the addition of exemplar essays as fixed reference points improved inter-rater agreement when assessing text structure, added evidence of validity, and resulted in scores with an increased generalisability.

In Chapter 3, the use of revision tests (RTs) as part of an assessment of writing was discussed. The validity and reliability of both a test with an existing multiple-choice (MC) format and a newly developed constructed-response (CR) format were evaluated. The MC test was found to cover a broad range of revision activities but lacked (face) validity because of the indirect response format. Conversion to a CR format added evidence of validity by offering a closer representation of the cognitive processes involved in text revision of the actual revision activity, but it led to a more restricted assessment of revision activities. These results led to the conclusion that a combination of both test formats would be best suited to ensure both validity and an adequate breadth of domain coverage.

The above results show that fairly successful efforts were made in this study to improve the validity and reliability of text quality scores. Still, the unwanted influences of writing tasks and raters and the difficulty to cover all facets of writing ability will—to some extent—persist when assessing writing. However, the effect of these issues on the validity and reliability of an assessment is dependent on the intended use of its test scores. Not all tests are designed to cover all aspects of the construct of writing, or to produce test scores that are used to make important decisions. That is, an assessment that only considers the end product of the writing process in order to produce a test score on text quality cannot claim to validly assess the entire multifaceted activity of producing a text, but it *can* validly assess a candidates ability to produce a text that meets certain predefined criteria. Likewise, a text quality score assigned by a single rater and based on a single writing task does not validly represent the range of all possible tasks and raters. Such a test is not suited for a high stakes examination, but it can be a useful tool to monitor progress in a classroom assessment.

In order to provide valid and reliable results for learning outcomes, a large-scale assessment aims to produce test scores that represent different facets of the writing process on the one hand, and that are generalisable across raters and tasks on the other hand. By presenting methods to assess revision ability and to improve generalisability of text quality scores, the present study has contributed to this goal.

### 5.1.2 Evaluating text complexity

In Chapter 4 of this study, a novel application of automated essay evaluation (AEE) was explored, namely the complexity evaluation of novice writers' texts. To our knowledge, this is the first study of its kind in the Dutch language. Given the aforementioned validity and reliability issues in evaluating text quality, an assessment of writing is likely to benefit from the use of a measure that is independent of specific task requirements, as opposed to text quality. For this reason, the use of text complexity features as indicators of writing ability is explored within this study. Whereas text quality is based on a *judgement* consciously constructed by a reader of a given text, text complexity is considered to be an *independent property* of a text, largely determined by objectively quantifiable linguistic features. A text quality judgement on the other hand, is dependent on specific evaluation criteria, determined by the intended communicative goal.

Recent technological advances enable automated analysis of several objectively extractable text features. This way, a fast and reliable analysis of linguistic features can be performed. In readability studies, such an analysis is employed in order to identify those textual features that influence text complexity—features that are thus considered to affect readability. In parallel, it might be considered that the production of complex text features demonstrates a high writing ability. Hence, linguistic complexity features are useful not only in readability research, but are potentially informative when assessing writing ability as well. Instead of using complexity features to evaluate the ability a *reader* needs in order to

*understand* a given text, those same features are used to assess the ability a *writer* needs in order to *produce* a text of a certain complexity (Figure 2). This way, evaluating text complexity features as indicators of writing ability might prove to be a useful addition to the evaluation of text quality by offering an objective and consistent metric to evaluate an array of essay features.

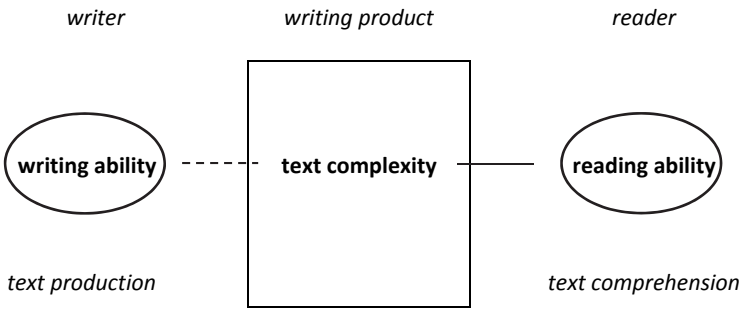



*Figure 2.* Text complexity, reading ability, and writing ability.

The study reported in Chapter 4 aimed to explore to what extent linguistic features that designate text complexity can be used as indicators of writing ability within automated essay evaluation (AEE). In general, text complexity was expected to increase with writing proficiency, since the ability to produce complex language is still developing in novice writers. In this study, the agreement between an array of complexity features and two indicators of writing ability—namely both grade level and human essay scores—was evaluated.

Results first of all revealed that characteristics of novice writing, such as limited text length and flaws in spelling and punctuation, unjustly influence complexity scores. Hence, manual pre-processing of the texts was needed in order to obtain valid results. Based on the analysis of 37 complexity measures, a selection of text complexity measures was found to be related to grade level and essay scores as indicators of writing ability (Table 1). These outcomes indicate that in novice writers, the ability to produce a complex text can be used as an indicator of writing ability. Furthermore, the outcomes highlight the need to develop methods to overcome unwanted influences on complexity on the one hand. and to further explore the relation between text complexity and the development in the writing ability of novice writers on the other hand. Nonetheless, by offering a consistent and fine-grained analysis of language use, AEE can readily be applied to characterise group differences within a large-scale assessment of writing.

Table 1. *Measures of Text Complexity and their Suggested Relation to Writing Ability*

| | **Measure** | **Suggested relation to writing ability** | |
|---|---|---|---|
| WORD | | | |
| Lexical complexity | Word length | + | Increases with ability |
| | Word complexity | + | Increases with ability |
| | Word frequency | - | Decreases with ability |
| Lexical richness | Adverbials | + | Increases with ability |
| | Type/token-ratio | - | Decreases with ability |
| | Prepositions | + | Increases with ability |
| SENTENCE | | | |
| Sentence complexity | Subordinate clauses | + | Increases with ability |
| | Dependency length | + | Increases with ability |
| | Clauses | - | Decreases with ability |
| | Sentence length | + | Increases with ability |
| TEXT | | | |
| Coherence | Argument overlap | - | Decreases with ability |
| | Referential pronouns | - | Decreases with ability |
| | Connectives | - | Decreases with ability |

## 5.2 Usability, future research, and practical applications

Notwithstanding the aforementioned results, the actual usability and validity of the methods investigated in this study are dependent on the intended use of the test scores—since different test purposes impose different requirements on these scores. For example, a selection test would only require a relative ordering of candidates to identify a given number of highest-scoring candidates, whereas a diagnostic assessment should provide detailed information on the performance of individual candidates. In this section, the applicability of the investigated methods within the context of a large-scale assessment in primary education is further evaluated. First, the preconditions of a large-scale assessment are specified, and the extent to which the applied methods meet these criteria is discussed, together with their applicability for novice writing (Section 5.2.1). Next, suggestions for future research and practical applications outside the context of a large-scale assessment are addressed (Section 5.2.2).

### 5.2.1 Assessing writing within a large-scale assessment in primary education

The present study was performed within the context of a national assessment on writing ability. This type of large-scale assessment is characterised by the objective to assess the quality of education by means of evaluating the learning outcomes. In order to accomplish this goal, national assessments differ from typical classroom assessments with respect to their scope, which has implications for the assessment design.

First, sound construct coverage is needed to provide a valid report on the learning outcomes of writing education by means of a large-scale assessment. In classroom practice, different aspects of writing are taught, monitored, and evaluated step-by-step, following a writing curriculum and taking into account the development of individual pupils. Hence, as part of a larger assessment programme, classroom assessment can suffice by covering a single aspect of writing at a time. On the other hand, a large-scale assessment aims at offering a complete evaluation of the writing construct by means of one test design. Hence, different writing tasks should be administered, ideally complemented by other test formats, in order to evaluate aspects of the writing process that are not reflected in the final writing product, such as the ability to revise a written text.

Second, whereas classroom assessment focuses on monitoring development in writing and identifying writing problems at an individual level, large-scale assessments evaluate the performance of an entire population, with results being disseminated to policymakers and teaching professionals, in order to improve the teaching practice nationwide. Therefore, instead of estimating the ability of individual writers, the ability of different groups of pupils is evaluated, enabling comparison of learning outcomes for pupils who differ in age, gender, socioeconomic status, or other background characteristics.

Construct representation

When assessing the construct of writing, different elements of the writing process, as well as the writing product, should be represented in order to ascertain construct coverage. Figure 3 presents a simplified model of the writing process in which four different components within the practice of producing a written text are related to elements of the writing product. As such, Figure 3 illustrates which specific components of the writing process are covered by each of the assessment approaches (AAA, AEE and RT) evaluated within this study.

The *planning* component—in which ideas are generated and content is organised—is covered by the newly developed AAA procedure. Within this method, the quality of written products was assessed by means of analytical questions together with a rating scale with anchor essays, covering the evaluation of content as well as text structure on the macro level.

The *translating* component concerns verbalising ideas into words, and organising these words into sentences and paragraphs. This component is represented in the AAA procedure by the evaluation of wording and text structure on the meso level (i.e. the composition of paragraphs and sentences). In addition to evaluating the *quality* of these aspects of the writing product, the *complexity* of paragraphs, sentences and words was evaluated by automated essay evaluation (AEE) in a separate study.

The *transcribing* of words in a correct manner was assessed in the AAA procedure, in which aspects of the correctness of language use were evaluated. Lastly, by administering a revision test (RT), a component of the writing process that is not reflected in the final written product is assessed, namely the *reviewing* and revising of text quality.

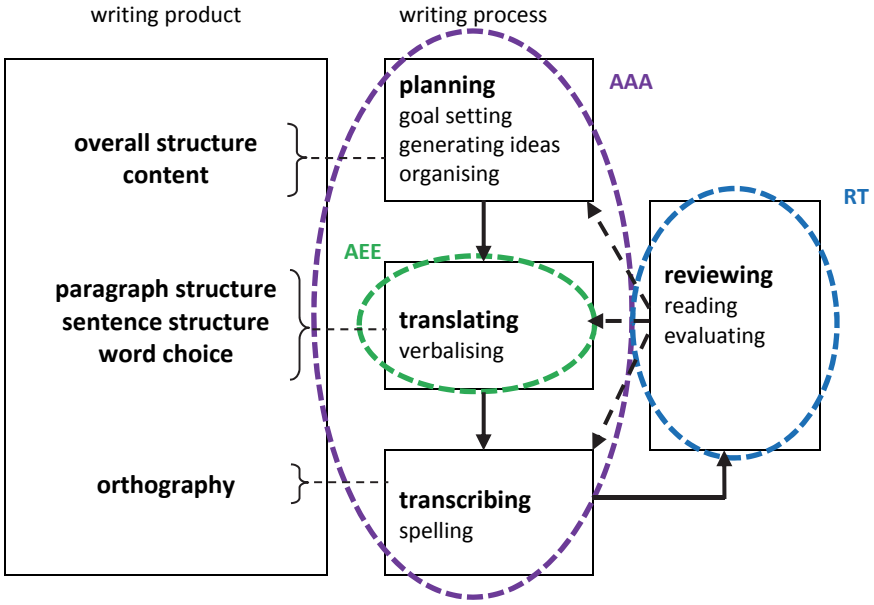



*Figure 3.* Elements of the writing product mapped to components of the writing process. (based on Flower & Hayes, 1981; Hayes, 2012; Van der Pool, 1995)

In Table 2, the coverage of the different elements of writing assessment is further substantiated for the three methods evaluated in this study. As such, Table 2 shows that all assessment methods contribute individually to the coverage of the writing construct. First, to assess the execution of the *planning* component of the writing process, the active construction of a writing product is required. To enhance generalisability, different types of writing tasks are to be administered to collect writing products of different genres. An analytical assessment enables a detailed analysis of these writing products, while anchor essays provide fixed reference points for raters, demanding evaluation of the written product as a whole. Moreover, the addition of anchor essays to an analytical assessment resulted in increased inter-rater reliability when assessing text structure within the present study. As such, the newly developed AAA procedure proved a valid and reliable method to evaluate the content and overall structure of the writing product, as well paragraph and sentence structure, and language use —reflecting planning, translating and transcribing skills (cf. Figure 3).

Second, the performance of novice writers is known to be restrained by the cognitively demanding process of *translating* thoughts into words (Kellogg, 2008) and *transcribing* words onto paper or screen (Hayes, 2012). Hence, an evaluation of the linguistic features produced is likely to provide discriminant information on the writing ability of different groups of novice writers. Although these linguistic features *can* be analysed by hand, this analysis would be time consuming and would require several thoroughly trained experts in order to yield valid and reliable results. Current natural language processing techniques offer the opportunity to perform this analysis efficiently and objectively by automating the evaluation of linguistic features. In the present study, the use of AEE was explored, and several measures of text complexity that appeared to be indicative of writing were identified.

Lastly, an alternative response format is needed when assessing the fourth element in the writing process (*reviewing*), since the ability to detect and correct flaws is not measurable by evaluating an end product. Instead, a writing product with flaws is provided, which is then corrected by the test taker. The present study has shown that while multiple-choice revision test items can be deployed to cover a broad range of revision activities, constructed-response revision items are needed to ensure (face) validity of the revision test. Hence, a combination of both constructed-response and multiple-choice items was proposed in Chapter 3.

Table 2. *Characteristics Per Assessment Method*

| | **Element of writing process** | | | | **Element of writing product** | | | | | **Text level** | | | **Evaluation criterion** | | **Response format** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | planning | translating | transcribing | reviewing | overall structure | content | sentence structure | word choice | orthography | macro | meso | micro | quality | complexity | productive writing | constructed response | multiple choice |
| AAA | x | x | x | | x | x | x | x | x | x | x | x | x | | x | | |
| RT | | | | x | | | x | x | x | | x | x | x | | | x | x |
| AEE | | x | x | | | | x | x | x | | x | x | | x | x | | |

Table 2 illustrates that together, the three methods evaluated in this study cover different aspects of the construct of writing as presented in this study (cf. Figure 1). However, this does not necessarily imply that all elements be administered in order to validly assess writing. Within an assessment, the validity of an instrument depends on the intended use of the test scores provided by this instrument. Hence, the administration of a revision test (RT) by means of an assessment of 'writing ability' would be invalid, while the use of an RT to evaluate students' skills in reviewing and improving text quality would be a valid interpretation of test scores. In other words, the three assessment methods evaluated in this study are separately deployable to evaluate different elements of the writing process, as well as different aspects of the writing product. On the other hand, the results presented in this study (and illustrated in Figure 1 and Table 2) suggest that when incorporating all elements (i.e., AAA, RT, and AEE) in a (large-scale) assessment of writing, all the main elements of the writing process as well as the writing product are addressed, hence supporting the validity of using this combination to evaluate the writing ability of novice writers.

Group assessment

Next to the aforementioned construct coverage, the three approaches of measuring writing ability that were evaluated in this study proved appropriate to compare groups of pupils within a large-scale assessment. First, the exemplars used within the AAA procedure provide informative illustrations of the differences in writing ability either across groups or within specific groups. These illustrations can be applied within the rating procedure, as well as when disseminating the results to the public, including teaching professionals.

Additionally, the evaluation of revision skills allows for an assessment of specific formulating problems, independent of the production of these elements in spontaneous writing. In other words, novice writers are known to be restricted by their formulating skills when writing (Kellogg, 2008) and hence are likely to avoid problematic formulations. By offering texts that are flawed and asking pupils within the national assessment to correct these flaws by either actively producing the correct formulation (constructed-response

format) or by choosing the correct alternative (multiple-choice format), insights into the abilities of different groups of pupils are given.

Lastly, the exploration of the use of AEE indicated that certain text complexity measures can be applied to evaluate formulating skills. However, analyses of outliers in text complexity scores revealed that the validity of text complexity measures was locally affected by the characteristics of novice writing. It means that some essays containing a relatively limited amount of words or a relatively large amount of flaws unjustly received high complexity scores. Since these effects are largely overcome when comparing large groups of pupils, AEE is specifically suited to be incorporated within a large-scale assessment. Moreover, by creating corpora of words or sentences and analysing random samples from these corpora, instead of essays, the effect of text length can presumably be eliminated, and a detailed analysis of the word use and sentence complexity for different groups of pupils is enabled.

Novice writing

When assessing writing products of novice writers, the specific properties of these products need to be taken into account. A characterisation of novice writing was given by Bereiter and Scardamalia (1987), who denoted the early stages of writing as 'knowledge telling', stating that novice writers are mainly concerned with getting their thoughts onto paper, and text production is hindered by the process of coding their thoughts into (correctly) written language. Hence, when considering the writing process as illustrated by Bereiter & Scardamalia (cf. Figure 1), the development in novice writing would mainly take place at the *translating* stage, a stance that is supported by Kellogg (2008), who stated that the writing ability of novice writers is still constrained by the process of translating thoughts into words. When both text quality and text complexity are considered, the methods evaluated in the present study proved suitable for the assessment of novice writers.

First, the fact that novice writers are constrained at the translation level will be reflected by limited textual fluency and flawed linguistic output. Indeed, analyses in the present study confirm that novice writers typically produce short texts with shortcomings, as illustrated in Figure 3. Given the aforementioned characteristics of novice writing, all approaches applied in this study can be considered suitable for the assessment of specific aspects of novice writing. To start with, the anchored analytical assessment (AAA) procedure benefits from the use of short texts, since the comparison of essay quality would probably be more complicated and less intuitive when longer texts are evaluated.

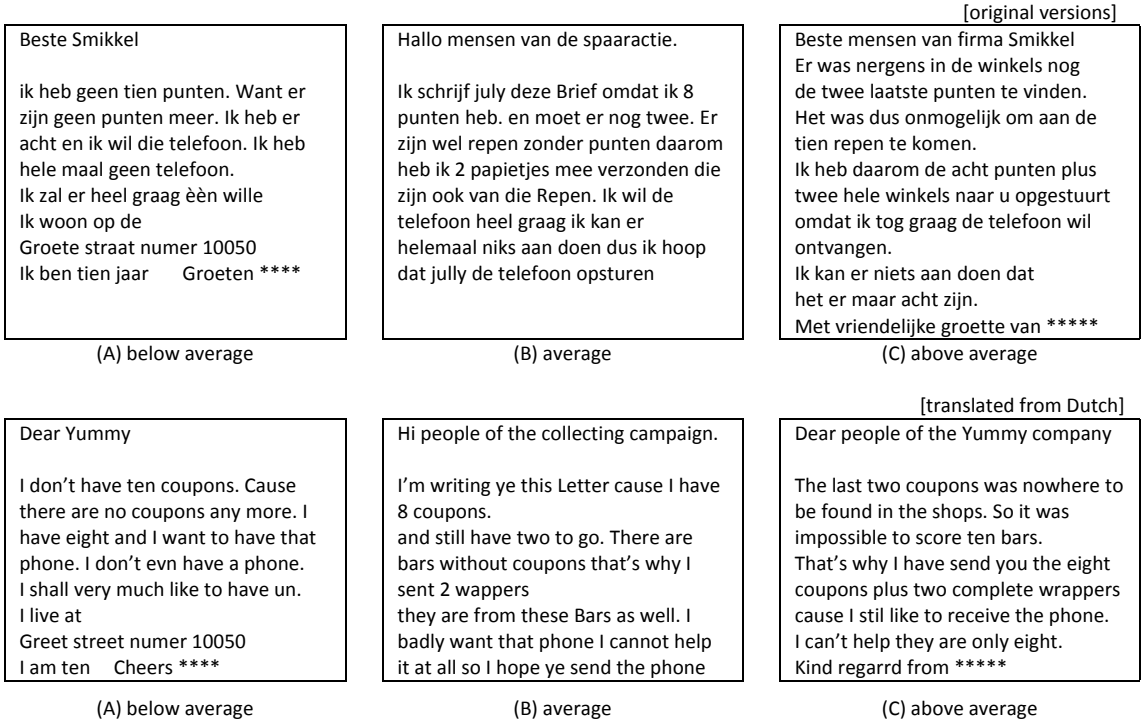



*Figure 4.* Different levels of writing ability in primary education.

Furthermore, revision tests were found suited to assess novice writers' ability to correct specific linguistic surface features. In novice writers, the process of translating thoughts into words is still cognitively demanding, hindering language production and resulting in flawed output (Bereiter & Scardamalia, 1987; Kellogg, 2008). Hence, the ability to detect and correct flaws when reviewing a given text is a discriminant factor when assessing novice writers, making revision tests a meaningful addition within an evaluation of novice writing. Automated essay evaluation (AEE) provides a method to objectively and efficiently evaluate the use of a large array of linguistic features, covering lexical richness, sentence complexity, and coherence. When supplemented with a pre-processing module, AEE is arguably suited to assess the correctness of word use, sentence construction, and use of cohesive devices.

Moreover, the fact that a number of text complexity features were found to be related to writing ability (cf. Table 1) indicates that AEE is especially suited to evaluate novice writers' texts, since text complexity is not expected to continually increase as writing ability develops. Rather, at a more advanced stage of proficiency, writers will start to adapt their writing to the intended reader and the purpose of the text, hence 'underperforming' if needed in terms of text complexity. In contrast, novice writers are still developing their skills to produce complex language, so that generally, it was found that even skilled novices did not yet apply their writing skills to reduce text complexity for the sake of clarity.

### 5.2.2 Future research and practical applications

The outcomes of this study give rise to a number of research questions to be answered in future research. First, it would be interesting to study the relations amongst the different approaches of writing assessment that were evaluated in the present study. Interlinking the results for each method will provide insights into the relations amongst different facets within writing ability, and possibly, different (developmental) patterns of strengths and weaknesses within writing can be identified.

The findings in Chapter 2 suggest that it would be useful to investigate the sources that account for differences in inter-rater agreement and generalisability amongst tasks and aspects of writing, in order to find ways to further improve both reliability and generalisability. Moreover, it would be noteworthy to examine the extent to which the anchored analytical assessment procedure can be used as a reliable and valid, stand-alone evaluation tool for different assessment purposes, including classroom assessment.

Furthermore, the cognitive processes involved in learning to review and revise a text would be a rewarding research topic. Gained insights into how to be a reader while writing a text can be employed to support the teaching and assessment of writing. The possibilities of designing a digital test format to assess revision ability should also be inves-tigated, in order to enable an objective assessment of a wider range of revision activities.

With respect to automated essay evaluation, future research should be aimed at acquiring a better understanding of the development of writing ability in general, and the relation to text complexity in particular. By further evaluating the validity of text complexity measures as indicators of writing ability, for example, by determining the correlations amongst (groups of) measures, insights will be gained into the specific textual features of high-quality texts. These findings should then be tested for consistency across different populations and writing contexts. Finally, to support language teachers and writing researchers, efforts should be made to disclose to the public the techniques underlying AEE methods.

In the near future, AEE is expected to aid the transition from (paper-based) summative assessment to computerised formative assessment. Within a digital writing environment, AEE can be applied to provide feedback on specific aspects of the writing product, offering writers a chance to adapt their texts—taking into account their writing goal and intended audience. Together with techniques to monitor the process of text production (e.g., keystroke logging, Leijten & Van Waes, 2013), digital writing offers information on the execution of the writing process, as well as the quality of the writing product. This information can be used by pupils to identify their strengths and weaknesses and to improve their writing, as well as by teachers to monitor their pupils' progress and to offer individualised feedback and instruction.

## 5.3 Conclusion

The present study aimed to improve the validity and reliability of writing assessment in primary education by evaluating different methods to assess both the quality and the complexity of novice writers' texts. An anchored analytical assessment (AAA) procedure provided raters with a fixed reference to be used when evaluating the text quality of writing products. A revision test (RT) was deployed to assess a writer's ability to review and revise the writing product in order to improve text quality—a part of the writing process that is not identifiable in the final writing product. Lastly, the evaluation of text complexity features was explored by means of automated essay evaluation (AEE). Together, these measures provided insight into the roles of text quality and text complexity as indicators of writing ability.

Figure 5 presents a schematic representation of writing assessment and the roles of text quality and text complexity therein. As a result of the writing process (A), a writing product (B) is produced, which is assigned a score (C). The writing product is considered a performance of writing ability (D) and is evaluated by taking both text quality (E) and text complexity (F) into account. The results of this study indicate that each of the three methods applied to assess text quality (RT, AAA) and text complexity (AEE) contributes to the reliability or validity of the eventual score on writing ability.

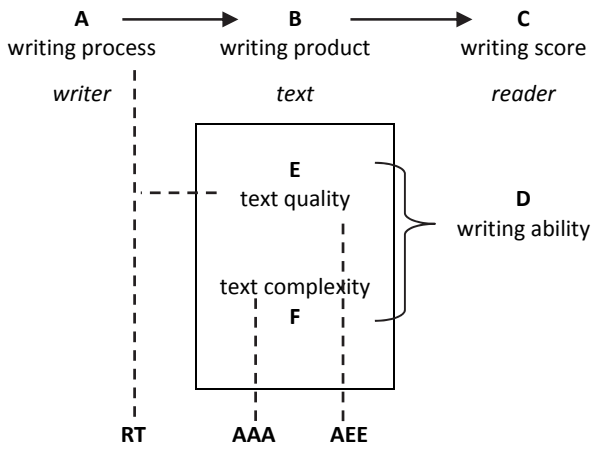



*Figure 5.* The schematic relations amongst text quality, text complexity, and writing ability.

Apart from exploring the use of text complexity features as indicators of writing ability, the present study provides insight into the relation between text complexity and text quality. In Chapter 4, the relation between text complexity features and text quality scores was evaluated. Results reported show that a selection of text complexity features correlate positively with text quality ratings, indicating that based on these features, a more complex text is considered of higher quality when compared to a less complex text. However, considering text complexity to be an aspect of text quality within an assessment of writing is only valid when the communicative goal of the text demands a certain complexity. That is, assigning a low quality rating to an absence note or birthday greeting because of its low

complexity, is considered an invalid use of text complexity scores. Furthermore, not all complexity features were found to correlate with text quality, suggesting a less straightforward relation between text complexity and text quality. Possibly, certain complexity features are indicative of the ability in early stages of writing development, whereas other features characterise the language production in later developmental stages.

Each of the three methods discussed in Chapters 2 to 4 of this study assesses a different aspect of the construct of writing ability. An anchored analytical assessment (AAA, Chapter 2) concerns the quality of a text, taking into account the specific text goal, while automated essay evaluation (AEE, Chapter 4) evaluates linguistic complexity features of the written product. Additionally, revision tests (RT, Chapter 3) focus on the writers' evaluation of text quality within the reviewing component of the writing process. The results presented suggest beneficial effects on both reliability and validity when applying these approaches in a writing assessment.

First, the addition of anchor essays to an analytical rating procedure was found to improve the agreement amongst raters, the validity of the writing scores, and the extent to which these text quality scores are generalisable across tasks and raters. Second, the use of a constructed-response format offered a valid method to assess the writers' ability to revise their texts in order to improve text quality. Finally, the automated evaluation of text complexity features proved promising as an objective metric to evaluate linguistic features that are indicative of the proficiency of novice writers. Overall, the evaluated methods were found to increase both the construct coverage and the consistency of the writing assessment, hence contributing to a valid and reliable assessment of writing ability in primary education.

## References


Abbott, R., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology*, *85*, 478–508.

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. Applied Psychological Measurement, 12, 117-128.

Almond, R., Quinlan, T. H., & Attali, Y. (2011). *Using timing logs to diagnose problems in writing performance.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Attali, Y. (2013). Validity and reliability of automated scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay scoring: current applications and future directions* (pp. 181-198). New York: Routledge.

Attali, Y., & Powers, D. (2008). *A developmental writing scale*. Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2. *Journal of Technology, Learning, and Assessment*, *4*(3). Retrieved from http://www.jtla.org

Bachman, L. F., & A S. Palmer (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bamberg, B. (1984). Assessing coherence: A reanalysis of essays written for the National Assessment of Educational Progress, 1969-1979. *Research in the Teaching of English, 18*(3), 305-319.

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance, *Assessment in Education: Principles, Policy & Practice, 18*(3), 279-293.

Barkaoui, K. (2007). Revision in second language writing: What teachers need to know. TESL Canada Journal/Revue TESL du Canada, 25(1).

Bartlett, E. J. (1982). Learning to revise: Some component processes. In M. Nystrand (Ed.), *What writers know: The language, process and structure of written discourse* (pp. 345-363). New York: Academic Press.

Bennett, R. E., & Bejar, I. I. (1998).Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–17.

Bennett, R. (1993). On the meanings of constructed response. In C. Ward & R. Bennett (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-28).

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *The Journal of Technology, Learning and Assessment*, *6*(1), 1–47.

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bereiter, C. (1980). Development in writing. In L. W. Gregg and E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 73-93). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bergh, H. van den, De Maeyer, S., Van Weijen, D. & Tillema, M. (2012). Generalizability of text quality scores. In E. van Steendam, M. Tillema, G. C. W. Rijlaarsdam & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices.* Leiden/Boston: Brill.

Bergh, H. van den (1990). Schrijfvaardigheid getoetst in het centraal schriftelijk eindexamen. [Assessing writing ability within the central written examinations.] *Levende Talen, 451*, 225-229.

Bergh, H. van den & Eiting, M. (1989). A method of estimating rater reliability. *Journal of Educational Measurement*, *26*, 29-40.

Bergh, H. van den (1988). Schrijven en schrijven is twee: Een onderzoek naar de samenhang tussen prestaties op verschillende schrijftaken. [Writing and writing is two. A study of the relationship between results on different writing tasks.] *Tijdschrift voor Onderwijsresearch, 13*(6), 311-324.

Bergh, H. van den, & Rijlaarsdam, G. (1986). Problemen met opstelbeoordeling? Een recept. [Issues with essay evaluation? A recipe.] *Levende Talen, 413*, 448–454.

Berninger, V., & Swanson, L. (1994). Modifying Hayes and Flower's model of skilled writing to explain beginning and developing writing. In E. Butterfield (Ed.) *Children's writing: toward a process theory of development of skilled writing* (pp. 57–81). Greenwich, CT: JAL.

Blok, H. & Hoeksema, J. B. (1984). Opstellen geschaald*. De constructie van beoordelingsschalen voor vijf schrijfopdrachten.* [Essays scaled. The construction of rating scales for five writing assignments.] Amsterdam: Stichting Centrum voor Onderwijsonderzoek van de Universiteit van Amsterdam.

Blood, I. (2012). *Automated essay scoring: A literature review*. Retrieved from http://www.tc.columbia.edu/tesolalwebjournal

Bouwer, R., Béguin, A., Sanders, T. & van den Bergh, H. (in press). Effect of genre on the generalizability of writing scores, *Language Testing,* 1–28.

Breetvelt, I. (1991). *Schrijfproces en Tekstkwaliteit. Een onderzoek naar het verband tussen Schrijfproces en Tekstkwaliteit bij leerlingen in het voortgezet onderwijs. [Writing process and text quality. A study on the relation between writing process and text quality in secondary education.]* Amsterdam: SCO.

Breland, H. (1999). *Exploration of an automated editing task as a GRE® writing measure.* GRE Board Report, No. 96-01R.

Breland, H., Camp, R., Jones, R. J., Morris, M. M. & Rock, D. A. (1987). *Assessing writing skill*. New York: College Entrance Examination Board.

Breland, H., & Gaynor, J. L. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement*, *76*, 119-128.

Brindley, G. (2001). Investigating rater consistency in competency-based language assessment. In G. B. C. Burrows, *Studies in immigrant English language assessment* (Vol. 2, pp. 59–80). Sydney: National Centre for English Language Teaching and Research, Macquarie University.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). Automated Evaluation of Discourse Coherence Quality in Essay Writing. In Shermis, M.D., & Burstein, J. (Eds.), *Handbook of Automated Essay Scoring: Current Applications and Future Directions* (pp. 267–280). New York: Routledge.

Burstein, J., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., . . . Wolff, S. (1998). *Computer analysis of essay content for automated score prediction: A prototype automated scoring system for GMAT analytical writing assessment essays.* ETS Research Report No. 98-15. Princeton, NJ: Educational Testing Service.

Camp, R. (2009). Changing the model for the direct assessment of writing. In B. Huot & P. O'Neill (Eds.). *Assessing writing. A critical sourcebook.* Urbana: Bedford, St. Martin's.

Campbell, J. R. (1999). *Cognitive processes elicited by multiple-choice and constructed-response questions on an assessment of reading comprehension.* (UMI No. 9938651)

Chanquoy, L. (2001). How to make it easier for children to revise their writing: A study of text revision from 3rd to 5th grades. *British Journal of Educational Psychology*, *71*, 15-41.

Chapelle. C., & Chung, Y. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing, 27*(3), 301–315.

Chapelle, C. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology* (Vol. 7). John Benjamins Publishing.

Chenoweth, A., & Hayes, J. (2003). The inner voice in writing. *Written Communication*, *20*, 99-118.

Cito, (1998-2010). *Entreetoetsen groep 5, 6 en 7*. [Entrance Tests grade 5, 6 and 7]. Arnhem: Cito.

Clauser B. E., Kane M. T., & Swanson D. B. (2002). Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, *15*(4), 413–432.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, *24* (4), 310–324.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement*. 2nd ed. (pp. 271–302). Washington, DC: American Council on Education.

Coffman, W. E. (1966). On the validity of essay test of achievement. *Journal of Educational Measurement, 3*, 151-156.

Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston, MA: Houghton Mifflin Company.

Cooper, P. L. (1984). *The assessment of writing ability: A review of research.* GRE Board Research Report GREB No. 82-15R, ETS Research Report 84-12.

COTAN (2010). *Beoordelingssysteem voor de kwaliteit van tests*. [Framework for the assessment of test quality.] Amsterdam: Nederlands Instituut van Psychologen.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: Wiley.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing* 7, 31–51.

Cushing Weigle, S. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Cushing Weigle, S. (2013). English language learners and automated scoring of essays: Critical considerations. *Assessing Writing*, *18*(1), 85–99.

Davis, B. G., Scriven, M. & Thomas, S. (1981). *The evaluation of composition instruction.* Point Reyes, CA: Edgepress.

Deane, P. (2013). On the relation between Automated Essay Scoring and Modern Views of the writing construct. *Assessing Writing, 18*, 7–24.

Deane, P. & Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. *Journal of Writing Research, 2*(2), 151-177.

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). Cognitive models of writing: Writing proficiency as a complex integrated skill. Educational Testing Service (ETS) Research Report 08-55. Princeton, NJ: ETS.

Elliot, S. (2003). Intellimetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Evers-Vermeul, J., & Sanders, T. (2009). The emergence of Dutch connectives: How cumulative cognitive complexity explains the order of acquisition. *Journal of Child Language*, *36*(4), 829–854.

Faigley, L., & Witte, S. (1981). Analyzing revision. *College Composition and Communication*, *32*(4), 400-414.

Flower, L. & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387.

Foltz, P. W., Landauer, T. K., & Laham, D. (1999). Automated essay scoring: Applications to educational technology. In *Proceedings of EdMedia '99*. Retrieved from ttp://www.psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html

Galbraith, D. (2009). Cognitive models of writing. *GFL - German as a Foreign Language, 2*(3), 7-22.

Gelderen, A. van, Oostdam, R. & Schooten, E. van (2011). Does foreign language writing benefit from increased lexical fluency? Evidence from a classroom experiment. *Language Learning, 61*, 281-321.

Gelderen, A. van. (1997). Elementary students' skills in revising: Integrating quantitative and qualitative analysis. *Written Communication*, *14*(3), 360-397.

Gelderen, A. van, & Blok, H. (1991). De praktijk van het stelonderwijs in de groepen 7 en 8 van de bassischool; observaties en interviews. [The practice of writing education in grades 7 and 8 in primary education: observations and interviews.] *Pedagogische Studiën 68*, 159–175.

Gelderen, A. van., & Blok, H. (1989). *Het stelonderwijs in de hoogste groepen van het basisonderwijs [Writing education in the upper classes of primary education].* Amsterdam: SCO.

Glopper, K. de (1985). Opstelkenmerken en opstelbeoordelingen. [Essay features and essay evaluations.] *Tijdschrift voor Taalbeheersing, 6*, 176-189.

Glopper, K. de (1988). *Schrijven beschreven. Inhoud, opbrengsten en achtergronden van het schrijfonderwijs in de eerste vier leerjaren van het voortgezet onderwijs.* [Writing described. Content, outputs and background of writing education in the first four years of secondary education.] (Diss. UvA). Den Haag: SVO.

Godshalk, F. I., Swineford, F. & Coffman, W. E. (1966). *The measurement of writing ability.* New York: College Entrance Examination Board.

Groenendijk, T,. Janssen, T., Rijlaarsdam, G., & Bergh, H. van den (2008). How do secondary school students write poetry? How creative writing processes relate to final products. *L1 – Educational Studies in Language and Literature*, *8*(3), 57-80.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hardison, C. M., & Sackett, P. R. (2008). Use of writing samples on standardized tests: Susceptibility to rule-based coaching and the resulting effects on score improvement. *Applied Measurement in Education, 21*(3), 227–252.

Hayes, J. R. (2012). Modelling and remodelling writing. *Written Communication 29*, 369.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Randell (Eds.), *The science of writing. Theories, methods, individual differences and applications.* (pp. 1-27). Mahwah, NJ: Lawrence Erlbaum.

Hayes, J. R., Flower, L. S., Schriver, K. A., Stratman, J. F., & Carey, L. (1987). Cognitive processes in revision. In S. Rosenberg (Ed.), *Advances in Applied Psycholinguistics, Vol. 2: Reading, Writing and Language Learnin*g (pp. 176-240). Cambridge: Cambridge University Press.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. Gregg & E. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 3–30). Hillsdale, NJ: Lawrence Erlbaum associates.

Hermans, P., Van der Schoot, F., Sluijter, C. & Verhelst, N. (2001). *Balans van depeiling beeldende vorming aan het einde van de basisschool 2. Uitkomsten van de tweede peiling in 1996.* [Report of the national assessment on visual arts at the end of primary education 2. Outcomes of the second survey in 1996.] Arnhem: Cito.

Heuvelmans (2011). *TiaPlus©.* M. & R. Department, Cito, Arnhem.

Hoeven, J. van der (1997). Children's composing: A study into the relationships between writing processes, text quality, and cognitive and linguistics skills. *Utrecht Studies in Language and Communication, 12*, Amsterdam: Rodopi.

Hohensinn, C., & Kubinger, K. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement*, *71*(4), 732-746.

Hunt, K. (1966). Recent measures in syntactic development. *Elementary English*, *43*, 732–739.

Huot, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication*, *47*, 549–566.

In'nami, Y., & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing, 26*, 219-244.

Inui, K., Tokunaga, T., & Tanaka, H. (1992). Text revision: a model and its implementation. In *Aspects of automated natural language generation*: *Proceedings of the 6th International Workshop on Natural Language Generation, Trento, Italy* (pp. 215-230). Berlin: Springer.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement*, (4th ed.) (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1-26.

Kellogg, R. T. (2006). A model of working memory in writing. In: C. M. Levy & S. Randsdell (Eds.), *The science of writing* (pp. 57-72). Mahwah, NH: Lawrence Erlbaum Associates.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*, 81-96.

Knoch, U., Read, J. & Von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*, 26-43.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? *Vocabulary Learning and Instruction, 1*(1), 60–69.

Krahmer, E., & van den Bosch, A. (2006). Punctuating Penelope. In *Liber amicorum Jaap Goedegebuure* (unpublished).

Krom, R., Gein, J. van de, Hoeven, J. van der, Schoot, F. van der, Verhelst, N., Veldhuijzen, N., & Hemker, B., (2004). *Balans van het schrijfonderwijs op de basisschool. Uitkomsten van de peilingen in 1999: halverwege en einde basisonderwijs en speciaal onderwijs. [Report of the national assessment on writing education in primary schools. Outcomes of the surveys in 1999: mid and end of primary education and special educational needs.]* Arnhem: Cito.

Kuhlemeier, H., Til, A. van, Hemker, B., Klijn, W. de, & Feenstra, H. (2013). *Balans van de schrijfvaardigheid in het basis- en speciaal onderwijs 2. Uitkomsten van de peiling in 2009 in groep 5, groep 8 en de eindgroep van het SBO. [Report of the national assessment on writing in primary and special education. Outcomes of the survey in 2009 for grade 3, grade 5 and final grade in special education.]* Arnhem: Cito.

Kuhlemeier, J.B. (1996). *Taalvaardigheid, taalactiviteiten en taalattitudes. Een validatiestudie.* [Language ability, language activities and language attitudes. A validation study.] (Diss. UvA). Arnhem: Cito.

Lee, Y., C. Gentle, & R. Kantor (2009). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics*, *31*(3), 391–417.

Lee, Y. W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater.* Princeton, NJ: Educational Testing Service.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using input log to analyze and visualize writing processes. *Written Communication, 30*(3), 358–392.

Levelt, S. (1989). *Speaking: From intention to articulation*. Cambridge, Mass.: MIT Press.

Linterman-Rygh, I. (1985). Connector density – an indicator of essay quality? *Text, 5*, 347-357.

Lissitz, R. & Hou, X. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology, 13*(3).

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33-68).

Loban, W. (1976). Language development: Kindergarten through grade twelve. Research Report No. 18. Urbana, IL: National Council of Teachers of English.

Lord, F. M., Novick, M. R. & Birnbaum, A. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McMillan, J. H. (2008). *Assessment essentials for standards-based education* (2nd ed.). Thousand Oaks: Corwin Press, SAGE Company.

McCutchen, D., & Perfetti, C. A. (1982). Coherence and connectedness in the development of discourse production. *Text*, *2*(1/3), 113–139.

McNamara, D., Crossley, S., & P. McCarthy (2010). Linguistic Features of Writing Quality. *Written Communication, 2011*(27), 57.

McNamara, T. (2000). *Language Testing.* Oxford: Oxford University Press.

McNamara, T. (1990). Item response theory and the validation of an ESP test for health professionals. *Language Testing, 7*(1), 52-76.

Melse, L. & Kuhlemeier, H. (2000). *Beoordeling van schrijfproducten met en zonder ankers. Is er verschil?* [Evaluating writing products with and without anchor essays. Is there a difference?] Publicaties Voortgezet Onderwijs. Arnhem: Cito.

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In C. Ward & R. Bennett (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 61-74).

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, D.C.: American Council on Education.

Meuffels, B. (1994). *De verguisde beoordelaar; Opstellen over opstelbeoordeling. [The maligned rater: Essays on essay evaluation].* Amsterdam: Thesis Publishers.

Milliano, I. de, Van Gelderen, A., Sleegers, P., Van Schooten, E. (2012). Het belang van motivatie en lesparticipatie in taal- en zaakvakken voor de schrijfontwikkeling van vmbo-leerlingen. [The importance of motivation and participation in class for the writing development of pre-vocational school students. In N. de Jong, K. Juffermans, M. Keijzer & L. Rasier. *Papers of the Anéla 2012 Applied Linguitstics Conference*. Delft: Eburon.

Mullis, I., Dossey, V. S., Campbell, J. A., Gentile, J. R., O'Sullivan, C. A., & Latham, A. S. (1994). *NAEP 1992 trends in academic progress*. Washington, D.C.: National Center for Education Statistics.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*(1), 1-18.

Page, E. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan.* 47(5), 238–243.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, *76*, 561–565.

Perfetti, C. & McCutchen, D. (1987). Schooled language competence: Linguistic abilities in reading and writing. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics, reading, writing and language learning,* Vol. 2 (pp. 105–141). New York: Cambridge University Press.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*(4), 651–6711.

Pollitt, A. (2004). *Let's stop marking exams.* Paper presented at the IAEA Conference, Philadelphia, June 2004.

Pool, E. van der (1995). *Writing as a conceptual process: A text analytical study of developmental aspects.* Dissertatie Universiteit van Tilburg.

Pullens, T., den Ouden, H., Herrlitz, W. & Bergh, H. van den (2013). Kan een meerkeuzetoets bijdragen aan het meten van schriftelijke taalvaardigheid? [Can a multiple-choice test contribute to the measurement of writing skill?] In: *Levende Talen Tijdschrift*, *14*(2), 31-41.

Ramineni, C. (2013). Validating automated essay scoring for online writing placement. *Assessing Writing*, *18*(1), 40–61.

Ramineni. C., & D. Williamson (2013). Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, *18*(1), 25–39.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language*, *14*, 201–209.

Rijlaarsdam, G., Van den Bergh, H. Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Van Steendam, E. & Raedts, M. (2012). Writing. In K. R. Harris, S. Graham & T. Urdan (Eds.), *APA educational psychology handbook: Application to learning and teaching* (Volume 3) (pp. 189-228). Washington, D.C.: American Psychological Association.

Rijlaarsdam, G. (1986). *Effecten van leerlingrespons op aspecten van stelvaardigheid.* [Effects of student response to aspects of writing ability.] SCO-rapport 88. Amsterdam: UvA.

Rodriguez, M. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct-writing assessments. *Applied Measurement in Education, 7*(2),159-170.

Sanders, T. & Spooren, W. (2007). Discourse and text structure. In D. Geeraerts, & J. Cuykens (Eds.), *Handbook of cognitive linguistics*, Oxford: Oxford University Press, 916-941.

Sanders, T. & Schilperoord, J. (2006). Text structure as a window on the cognition of writing; How text analysis provides insights in writing products and writing processes. In C. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research*, New York: Guilford Press, 386-402.

Sanders. T. & Van Wijk, C. (1996a). PISA - A procedure for analyzing the structure of explanatory texts. *Text, 16*, 91-132.

Sanders, T. & Van Wijk, C. (1996b). Text analysis as a research tool: How hierarchical text structure contributes to the understanding of conceptual processes in writing. In M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 251-269). Mahwah NJ: Erlbaum.

Sanders, T., Janssen, D., Van der Pool, E., Schilperoord, J. & Van Wijk, C. (1996). Hierarchical text structure in written products and writing processes. In G. Rijlaarsdam, H. van den Bergh & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 473-492). Amsterdam: Amsterdam University Press.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15*, 1–35.

Schoonen, R. (2011). How language ability is assessed. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. II) (pp. 701-716). Routledge.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modelling. *Language Testing, 22*, 1-30.

Schoonen, R., Vergeer, M. & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*, 157-84.

Schoonen, R. & De Glopper, C. M. (1992). Toetsing van schrijfvaardigheid: problemen en mogelijkheden. *Levende Talen, 470*, 187-195.

Schoonen, R. (1991). *De evaluatie van schrijfvaardigheidsmetingen. Een empirische studie naar betrouwbaarheid, validiteit en bruikbaarheid van schrijfvaardigheidsmetingen in de achtste groep van het basisonderwijs.* Amsterdam: Universiteit van Amsterdam/SCO.

Schooten E. van & Glopper, K. de (1990). De validiteit van meerkeuze-instrumenten voor het meten van schrijfvaardigheid. *Tijdschrift voor Taalbeheersing*, *12*, 93-110.

Schwartz, J. A. & Collins, M. L. (1995). *Improving the reliability of a direct writing skills assessment.* Paper presented at the Nineteenth Annual PIMAAC Conference. New Orleans, LA.

Shaw, S. D. & Weir, C. J. (2007). *Studies in language testing 26: Examining Writing. Research and practice in assessing second language writing.* Cambridge: Cambridge University Press.

Shermis, M., Burstein, J., & Apel-Bursky, S. (2013). Introduction to automated essay evaluation. In J. Burstein & M. Shermis (Eds.), *Handbook of automated essay evaluation: current applications and new directions* (pp. 1–15). New York: Routledge.

Shermis, M., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed., pp. 20–26). Oxford, UK: Elsevier.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: current applications and new directions* (pp. 153-180). New York: Routledge.

Sijtstra, J. (Ed.) (1997). *Balans van het taalonderwijs aan het einde van de basisschool 2. Uitkomsten van de tweede taalpeiling einde basisonderwijs*. [Report of the language education at the end of primary education 2. Outcomes of the second language survey end primary education.] Arnhem: Cito.

Silva, M. L., Sánchez Abchi, V., & Borzone, A. (2010). Subordinated clauses usage and assessment of syntactic maturity: A comparison of oral and written retellings in beginning writers. *Journal of Writing Research, 2*(1), 47–64.

Sluijter, C., Verhelst, N. & Hermans, P. (1999). *De ontwikkeling van een procedure voor het betrouwbaar en valide meten van tekenvaardigheid.* [The development of a procedure for a reliable and valid measurement of drawing skill.] Paper presented at the Onderwijs Research Dagen [Educational Research Days] 1999.

Snow, R. E. (1993). Construct validity and constructed-response tests. Construction versus choice. In *Cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60).

Sommers, N. (1980). Revision strategies of student writers and experienced writers. *College Composition and Communication*, *31*, 378-387.

Staphorsius, G. (1994). *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument.* Enschede: University of Twente.

Stevenson, M., Schoonen, R., & Glopper, K. de (2006). Revising in two languages: A multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing, 15*(3), 201–233.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. E., Reed, M. S., . . . White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*(3), 492–506.

Til, A. van, van Weerden, J., Hemker, B., & Keune, K. (2013). *Balans van de taalverzorging en grammatica in het basis- en speciaal basisonderwijs. Uitkomsten van de peiling in 2009 in groep 5, groep 8 en de eindgroep van het SBO.* Cito.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In C. Ward & R. Bennett (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 29-44).

Weijen, D. van (2008). *Writing process, text quality, and task effects. Empirical studies in first and second language writing.* LOT: Utrecht.

Wesdorp, H. (1974). Het meten van de produktief-schriftelijke taalvaardigheid*. Directe en indirecte methoden: 'opstelbeoordeling' versus 'schrijfvaardigheidstoetsing'. [Measuring productive written language ability. Direct and indirect methods: 'essay evaluation' versus 'assessment of writing ability']* Purmerend: Muusses.

Williamson, D. (2013). Probable cause: developing warrants for automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: current applications and new directions* (pp. 153-180).New York: Routledge.

Williamson, D., Xi, X., & Breyer, J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13.

Witte, S., & Faigley, L. (1981). Coherence, cohesion, and writing quality: College composition and communication. *Language Studies and Composin*g, *32*(2). 189–204.

Wolcott, W., & Legg, S. M. (1998). *An overview of writing assessment: Theory, research and practice.* Urbana, IL: National Council of Teachers of English.

Yang, Y., Buckendahl, C., Juszkiewicz, P., & Bhola, D. (2002). A review of strategies for validating computer-automated scoring, Applied Measurement in Education, 15(4), 391–412.

Zwaan, R. A. & Rapp, D. N. (2006). Discourse comprehension. In M. A. Gernsbacher & M. J. Traxler (Eds.), *Handbook of psycholinguistics* (pp. 725-764). San Diego, CA: Elsevier.

Zwarts, M. (Ed.) (1990). *Balans van het taalonderwijs aan het einde van de basisschool. Uitkomsten van de eerste taalpeiling einde basisonderwijs.* [Report of the language education at the end of primary education 2. Outcomes of the first language survey end primary education.] Arnhem: Cito.

**Applications**

**T-Scan**

Kraf, R., van der Sloot, K., Pander Maat, H., van den Bosch, A., & Kleijn, S. (2013). *T-Scan*. Retrieved from http://languagelink.let.uu.nl/tscan

**Frog**

Bosch, A. van den, Busser, G. J., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In F. van Eynde, P. Dirix, I. Schuurman, & V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting* (pp. 99–114). Leuven, Belgium. Retrieved from http://ilk.uvt.nl/frog

**Alpino**

Noord, G. van (2013). *Alpino*. Retrieved from http://www.let.rug.nl/~vannoord/alp/Alpino

## Acknowledgements

Writing this thesis about writing has been challenging, intense, and frustrating at times. But most of all, it was an interesting process that has been educational in many ways. Producing a dissertation is a writing task that requires many solitary hours of writing (and rewriting...), but at the same time, it is a task that cannot be completed without the support of others. Many people have made valuable contributions throughout my writing process. I would hereby like to thank dearly all of you who have contributed to the realisation of this piece of writing.

In addition, some of you deserve a special mention in this section. First and foremost, my much-appreciated supervisors have contributed greatly throughout the course of my research project. I enjoyed our animated and valuable sessions together, which were also a good opportunity to catch up on the latest football news. Theo, thank you for enthusiastically joining my project and for defending its existence. You have been of great support during the process, and your comments were valuable in many ways. I admire your optimism and engagement throughout the project, especially over the past few months. Ted, your critical eye motivated me to keep challenging myself, and I have learned a lot from our talks. Whether in person or from a distance, you managed to inspire me and to make significant contributions to the process.

My PhD project was supported by Cito, and I am thankful for the opportunity I have been given to conduct this research. My thanks go out to two persons in particular. Gerrit, you understood I wanted to be a part-time researcher even before I did myself, and I am very grateful for the opportunity you created for me. And Alex, thank you for your moral support during the final stages of my project and for helping me find the peace of mind I needed to finish this piece.

Next, I feel privileged to have been able to work with some excellent researchers at Utrecht University during this project. Huub, while your knowledge and skills are intimidating, your supportive and helpful approach was very much appreciated. Henk, thank you for helping me interpret my T-Scan results and for your thought-provoking comments. Rogier, thanks for introducing me to T-Scan and for always being willing to answer my rookie questions. Suzanne, it has been a real pleasure to be your occasional roommate at Trans. Marloes, thank you for making me notice and even visit the carillon of the Dom Tower. Monica, Gerdien and all other lunch mates, thank you for making me feel welcome at Trans. Also, I am grateful to have been a part of the ever-growing group of ‘schrijfaio’s’. A special thanks goes out to Renske, who helped me find my way in the wondrous world of writing assessment. Thank you for agreeing to support me as a paranymph.

At Cito, I was supported in many ways by many colleagues. Riny, you have been of great help by providing me with additional data. Bas, you were always willing to perform yet another analysis. Hans and Karin, your comments on my writing were insightful and helped improved the quality greatly. Erna, thank you for enriching my dissertation with your endearing and clever translations of the pupils' writings. I also thank my fellow PhD students at POK/RCEC for their collegiality. And I especially thank Karen, who ingeniously analysed (and reanalysed) my data. Thank you for your moral support and for agreeing to be my paranymph.

Of course, writing a dissertation is not solely a professional affair. I am fortunate to have received support from many loved ones. To my dear parents: thank you for your upbringing that stimulated my curiosity, creativity and self-dependence, and for your support throughout. I also thank my beloved sisters Stefanie and Heleen: the 'sisterdays' we had together offered me much-needed moments of distraction and relaxation. Gerarda, I am glad to have more time now to travel to Groningen and take nice walks together. Helma, I hope we will be able to commemorate our time in primary school for many more years. Thank you for creating a wonderful cover for this book.

Even after writing a dissertation, much remains to be explored in future research. However, one conclusion I am able to draw without the usual academic reservations, is that without the continuous support of my loving husband Albert, I would not have been able to complete this piece. Thank you for not complaining about me working all those nights and weekends, for taking care of me in the meantime, for harassing me with your dreaded work schemes, and for meticulously arranging the most amazing travels that have broadened my horizons vastly. But most of all, I thank you for your belief in me.

Hiske Feenstra

Dieren, July 2014

## Summary


### Introduction

Writing is a complex ability, and measuring writing ability is a notoriously complex task. Producing a written text involves executing a set of cognitive activities, which together constitute the concept of writing ability. Within a writing assessment, the quality of the text produced is then evaluated by one or more raters. The assessment of writing ability is complicated by the multi-faceted nature of this productive language ability on one hand, and the difficulty of evaluating writing performances on the other hand. In this dissertation, different methods to assess writing ability are evaluated. The goal of this study is to improve the evaluation of writing ability within a large-scale assessment in primary education by addressing the following research question: *How can the writing ability of novice writers be assessed reliably and validly?*

To answer this question, three methods to assess writing ability are evaluated, each in a separate chapter. This study combines the traditional approach of assessing writing by means of evaluating *text quality* (Chapter 2 and Chapter 3), with a more innovative approach which explores the use of language technology to evaluate *text complexity* (Chapter 4). Together, these chapters aim to explore how a writing assessment can be designed such that it is suited for assessing different aspects of writing ability in primary education in a valid and reliable manner.

In **Chapter 1**, an introduction on the construct of writing ability and its assessment is presented. Producing a written text involves a series of activities, starting with the mental processes of generating, organising, and structuring ideas, and translating thoughts into words. Next, specific motor skills are needed to produce letters in either handwriting or typewriting, Finally, the writing process involves monitoring and editing the text thus far produced.

These different components of producing a text illustrate the multi-facetedness of writing. This feature complicates the design of an assessment which aims to rightly reflect writing ability, and thus threatens the validity of the assessment. In addition, the various components of a writing assessment—such as the writing task, rating procedure, and rater characteristics—are potential sources of construct-irrelevant variance (Messick, 1989). In other words, the characteristics of writing assessment may lead to a difference in writing scores that is not solely caused by the difference in ability between two candidates, but rather by a difference in the specific rater or writing task assigned to the candidate. This affects both the reliability and the validity of writing assessment (see Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). In this study, alterations in the Dutch national assessment of writing are evaluated, with the aim of improving the reliability and validity of the assessment.

**Results**

Within a writing assessment, a candidate typically writes a text in response to a specific writing task, after which the text produced is evaluated by one of more raters. The reliability and validity of writing assessments are affected by the fact that raters disagree on the quality of a text, on one hand, and the question to what extent a rating procedure truly represents the construct of writing ability, on the other hand. In **Chapter 2**, a newly developed assessment procedure, in which a rating scale with anchor essays is added to an analytical rating procedure, is evaluated. Having exemplar essays as a fixed reference (i.e., an 'anchor') is expected to help raters agree on the quality of the essays, resulting in a higher inter-rater agreement compared with that achieved a solely through analytical evaluation. In addition, the anchored analytical assessment (AAA) is predicted to diminish the effect of unwanted sources of variance—such as the task or rater—on the generated scores, hence resulting in highly generalisable writing scores.

In an experimental study, the inter-rater agreement, generalisability, and construct validity of the AAA-procedure is addressed. Essays collected in grade 3 (age 8) through grade 6 (age 12) were evaluated on the aspects content, structure and correctness. A comparison of the AAA-procedure with a solely analytical assessment procedure shows that the addition of anchor essays provides for significantly higher agreement between raters for the assessment of text structure. Furthermore, the results of a multiple regression analysis and a generalisability study indicate that the use of anchor essays enhances the construct validity of a writing assessment. These findings are supported by results from an analysis of convergent and divergent validity. In addition, a qualitative evaluation of the AAA-procedure shows that the use of anchor essays is found to be an operable method within a large-scale assessment.

**Chapter 3** discusses the use of revision tests within an assessment of writing. A skilled writer is constantly reviewing and rewriting the text produced thus far—with the target reader in mind—in order to improve the quality of the text (Bereiter & Scaradamalia, 1987; Kellogg, 2008). The ability to revise a written text is therefore an important skill to acquire and a relevant skill to monitor within an assessment. The aim of the study presented in Chapter 3 is to investigate how a revision test can be validly and reliably incorporated into a large-scale assessment of writing ability in primary education. First, the validity of an existing standardised, multiple-choice revision test is evaluated. Second, a constructed-response version of this test is piloted, and its validity and test characteristics are compared with those of the existing multiple-choice version.

A constructed-response task offers an authentic representation of the process of actively editing a text at the surface level. Furthermore, the analyses of test characteristics show that the reliability and discriminative power of the constructed-response test format are satisfactory. However, the domain coverage of (paper-based) constructed-response revision tasks is limited, since not all elements of the revision process are objectively assessable in an open-ended format. By contrast, a multiple-choice format is considered less

authentic because of its passivity, but offers a broader domain coverage since a larger array of textual features can be assessed (e.g., paragraph order). The content analyses of both formats of the revision tests however, show a clear focus on formulating skills. Therefore, in their present form these tests cannot be considered a valid representation of all components of revision ability. Instead, claiming that the evaluated tests measure the ability of pupils to perform revision activities on the surface level (e.g., language use) seems justifiable. Consequently, the results of this study suggest the construction of a revision test which combines both constructed-response and multiple-choice items, in order to accomplish a broader domain coverage and thus improve the assessment of revision ability.

In **Chapter 4**, the use of automated essay evaluation (AEE) is explored to gain knowledge on the relation between text complexity measures and writing ability, as well as the possibilities of using these measures as a part of a writing assessment in primary education. Given the aforementioned validity and reliability issues in evaluating text quality, text complexity is used in this explorative study to provide a measure of writing ability which is independent of the context required by the writing task, as opposed to text quality. T-Scan, a program for the automated evaluation of text complexity in Dutch (Kraf, Van der Sloot, Pander Maat, Van den Bosch & Kleijn, 2013) is used to extract the relevant textual features. The results first of all reveal that the validity of numerous text complexity measures is influenced by the fact that writing products in primary education are typically flawed, as is illustrated in Figure 1. These findings highlight the need to develop a pre-processing module (Pre-Scan), in which spelling and punctuation errors are detected and subsequently corrected.

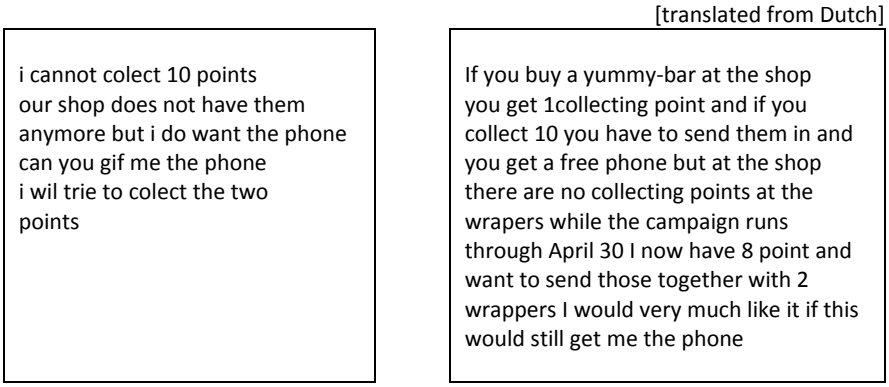



*Figure 1*. Examples of flawed essays.

In addition, 13 measures of text complexity that appear to be indicative of writing ability were selected, based on quantitative analysis of agreement between 37 complexity measures from T-Scan and both grade level and human essay scores. The qualitative evaluation of their validity and interpretability as measures of writing ability highlights the need to adjust several measures of text complexity, so that they are specifically suited for evaluating the writing ability of novice writers. Furthermore, the need to limit the influence of factors which unjustly influence complexity scores (e.g., text length) has become evident.

In conclusion, the results of this study indicate that in its current state, AEE is capable of offering a consistent metric and a fine-grained analysis of quality traits related to the linguistic surface characteristics of essays. The results of a qualitative study also highlight the need for further research to substantiate the usability of AEE within primary education. Nonetheless, AEE can be readily applied to characterise group differences across particular traits in large-scale tests. The evaluation of text complexity can therefore support writing education by offering an objective and informative method to complement the assessment of writing ability.

Lastly, **Chapter 5** provides an overall discussion of the results presented in previous chapters, as well as recommendations on the evaluation of writing ability within a large-scale assessment in primary education. The results presented in this study imply that the methods applied within this study cover different stages in the practice of producing a text (i.e., the writing *process*) by evaluating different parts of the results of this process (i.e., the writing *product*). Moreover, the present study combines two approaches to evaluate writing ability by addressing both the *quality* and the *complexity* of written texts. This way, information is gathered on the relation between specific features of a written product, on one hand, and the ability of its writer, on the other.

**Conclusion**

This study aimed to improve the validity and reliability of writing assessment in primary education by evaluating different methods to assess both the *quality* and the *complexity* of novice writers' texts. Figure 2 presents a schematic representation of writing assessment and the roles of text quality and text complexity therein. As a result of the writing process (A), a writing product (B) is produced, which is assigned a score (C). The writing product is considered a performance of writing ability (D) and is evaluated with both text quality (E) and text complexity (F) taken into account.

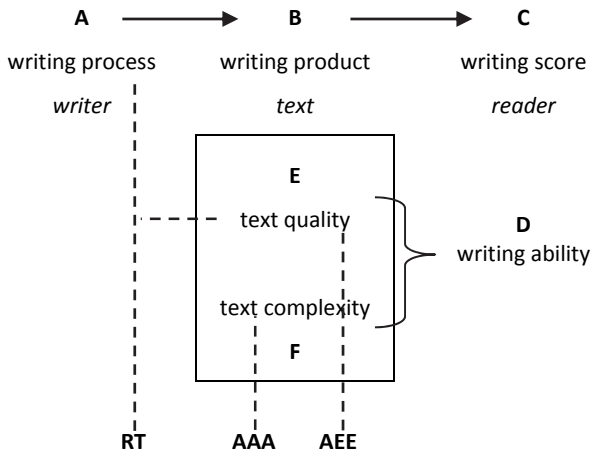



*Figure 2.* The schematic relations amongst text quality, text complexity, and writing ability.

Each of the three methods discussed in this study assesses a different aspect of the construct of writing ability, as is illustrated in Figure 2. Revision tests (RT, Chapter 3) focus on the writers' evaluation of text quality within the reviewing component of the writing process. An anchored analytical assessment (AAA, Chapter 2) concerns the quality of a text, with the specific goal of the text considered. Lastly, automated essay evaluation (AEE, Chapter 4) evaluates linguistic complexity features of the written product.

The results presented in this dissertation indicate several beneficial effects on both reliability and validity when the abovementioned approaches are applied in a writing assessment. First, the addition of anchor essays to an analytical rating procedure was found to improve the agreement amongst raters, the validity of the writing scores, and the extent to which these text quality scores are generalisable across tasks and raters. Second, the use of a constructed-response format offered a valid method to assess the writers' ability to revise their texts, so that text quality is improved. Finally, the automated evaluation of text complexity features proved promising as an objective metric to evaluate linguistic features which are indicative of the proficiency of novice writers. Overall, the methods evaluated within this study were found to increase both the construct coverage and the consistency of the writing assessment. Therefore, these methods contribute to a valid and reliable assessment of writing ability in primary education.

**References**

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bergh, H. van den, De Maeyer, S., van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. van Steendam, M. Tillema, G. C. W. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices.* Leiden/Boston: Brill.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1-26.

Kraf, R., van der Sloot, K., Pander Maat, H., van den Bosch, A., & Kleijn, S. (2013). *T-Scan*. Retrieved from http://languagelink.let.uu.nl/tscan

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*. (3rd ed., pp. 13–104). Washington, DC: American Council on Education.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modelling. *Language Testing, 22*, 1-30.

## Samenvatting

### Introductie

Schrijven is een complexe vaardigheid, en het meten ervan is een notoir lastige taak. Tijdens het produceren van een tekst voert een schrijver een verzameling van cognitieve activiteiten uit, die samen het construct schrijfvaardigheid vormen. Binnen een schrijfvaardigheids-meting wordt de kwaliteit van de geschreven tekst vervolgens door een of meerdere beoordelaars geëvalueerd. Het meten van schrijfvaardigheid wordt zowel bemoeilijkt door de gelaagdheid van de te meten vaardigheid, als door complicaties bij het objectief beoordelen van schrijfprestaties. In dit proefschrift worden verschillende methodes geëvalueerd om schrijfvaardigheid in het primair onderwijs te meten. Het doel van deze studie is om de schrijfvaardigheidsmeting binnen grootschalig (peilings)onderzoek te verbeteren door de volgende onderzoeksvraag te stellen: *Op welke wijze kan de schrijfvaardigheid van beginnende schrijvers betrouwbaar en valide getoetst worden?*

Om bovenstaande vraag te beantwoorden, worden in deze studie verschillende methoden van schrijfvaardigheidstoetsing geëvalueerd, elk in een apart hoofdstuk. Hierbij wordt ten eerste de traditionele benadering van schrijfvaardigheidstoetsing toegepast, namelijk het beoordelen van tekst*kwaliteit* (Hoofdstuk 2 en Hoofdstuk 3). Daarnaast wordt een meer innovatieve benadering gebruikt, waarin het gebruik van taaltechnologie bij het evalueren van tekst*complexiteit* wordt verkend (Hoofdstuk 4).

**Hoofdstuk 1** geeft een introductie op het construct schrijfvaardigheid en op de toetsing van deze complexe vaardigheid. Het schrijven van een tekst bestaat uit een reeks van activiteiten, te beginnen met een aantal mentale processen: het genereren, organiseren en structureren van ideeën, en het vertalen van deze ideeën naar concrete woorden. Daarna worden specifieke motorische vaardigheden ingezet om—handgeschreven of getypte—letters te produceren. Tot slot is een schrijver constant bezig om de tot dan toe geproduceerde tekst te evalueren en indien nodig te reviseren.

Bovengenoemde uiteenlopende activiteiten illustreren de gelaagdheid van schrijf-vaardigheid. Deze gelaagdheid bemoeilijkt het ontwerpen van een schrijftoets die recht doet aan alle componenten van schrijfvaardigheid, en vormt daarmee een bedreiging voor de validiteit van de toets. Daarnaast zijn de verschillende onderdelen waaruit een schrijfvaardigheidsmeting bestaat—zoals schrijftaak, beoordelaars, en beoordelings-procedure—potentiele bronnen van construct-irrelevante variantie (Messick, 1989). Met andere woorden: de eigenschappen van een schrijftoets kunnen ervoor zorgen dat een verschil in schrijfscores tussen twee leerlingen deels veroorzaakt wordt door bijvoorbeeld een verschil in toegewezen schrijftaak of beoordelaar, in plaats van door een verschil in vaardigheid. Hierdoor worden zowel de betrouwbaarheid als de validiteit van de toets negatief beïnvloed (zie Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). In deze studie worden wijzigingen in het toetsontwerp van de nationale schrijfpeiling onderzocht, met als doel om de betrouwbaarheid en validiteit van deze meting te verbeteren.

**Resultaten**

In een meting van schrijfvaardigheid schrijft een leerling over het algemeen een tekst naar aanleiding van een specifieke schrijfopdracht, waarna de geschreven tekst door een of meerdere beoordelaars beoordeeld wordt op tekstkwaliteit. De betrouwbaarheid en validiteit van deze wijze van schrijfvaardigheidstoetsing wordt enerzijds beïnvloed door het feit dat beoordelaars vaak verschillend oordelen over de kwaliteit van eenzelfde tekst, en anderzijds door de vraag tot op welke hoogte een gegeven oordeel het construct schrijfvaardigheid juist representeert. In **Hoofdstuk 2** wordt een nieuw ontwikkelde beoordelingsmethode geëvalueerd, waarin een beoordelingsschaal met voorbeeldopstellen is toegevoegd aan een analytische vragenlijst. Het aanbieden van voorbeeldopstellen (of 'ankers') als vast referentiepunt zal er naar verwachting voor zorgen dat beoordelaars het beter eens zijn over de kwaliteit van de te beoordelen opstellen. Hierdoor zal de interbeoordelaarsovereenstemming (een maat voor de betrouwbaarheid) van een beoordelingsmethode met ankers naar verwachting hoger zijn dan wanneer alleen een analytische beoordeling wordt toegepast. Bovendien zal de toevoeging van ankers er naar verwachting voor zorgen dat bronnen van ongewenste variantie—zoals de schrijftaak en de beoordelaar—minder effect hebben op de schrijfscores, wat de generaliseerbaarheid van de scores ten goede zal komen.

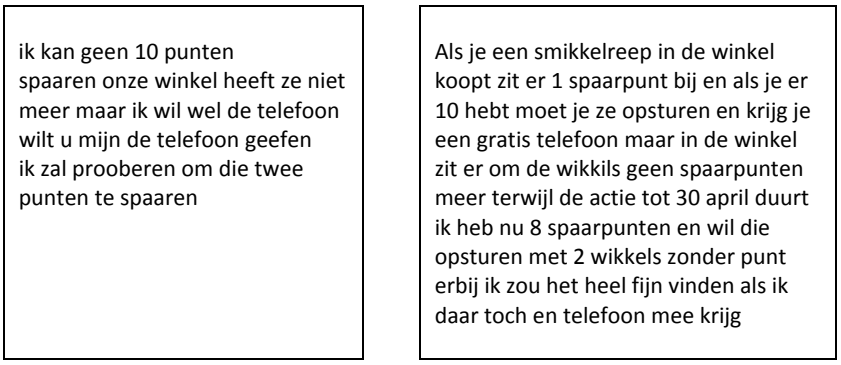
In een experimentele studie zijn interbeoordelaarsovereenstemming, generaliseerbaarheid en constructvaliditeit van de nieuw ontwikkelde analytische beoordeling met ankeropstellen (*anchored analytical assessment* [AAA]) onderzocht. In de groepen 5 tot en met 8 van het primair onderwijs zijn opstellen verzameld die door beoordelaars geëvalueerd werden op de aspecten 'inhoud', 'structuur' en 'taalverzorging'. Een vergelijking tussen de AAA-methode en de bestaande analytische methode laat zien dat de toevoeging van ankeropstellen zorgt voor een significant hogere betrouwbaarheid bij het beoordelen van structuurelementen in een tekst. Daarnaast laten de resultaten van een meervoudige regressieanalyse en een generaliseerbaarheidsstudie zien dat het gebruik van ankeropstellen bijdraagt aan de constructvaliditeit van een schrijfvaardigheidsmeting. De resultaten van een analyse van de convergente en divergente validiteit ondersteunen deze uitkomst. Tot slot blijkt uit een kwalitatieve evaluatie van de beoordelingsprocedure dat het gebruik van ankeropstellen een bruikbare methode is om in te zetten binnen een grootschalige toetsafname.

In **Hoofdstuk 3** wordt het gebruik van revisietoetsen (*revision tests* [RT]) binnen een schrijfvaardigheidsmeting besproken. Tijdens het schrijven van een tekst zal een ervaren schrijver de tot dan toe geproduceerde tekst constant evalueren en waar nodig reviseren om zo de kwaliteit van de tekst te verhogen—daarbij de beoogde lezer in gedachte nemend (Bereiter & Scaradamalia, 1987; Kellogg, 2008). Het kunnen reviseren van een tekst is daarmee een belangrijke vaardigheid om te verwerven, en een relevante vaardigheid om te toetsen. Het doel van de studie waarover gerapporteerd wordt in Hoofdstuk 3 is om te onderzoeken hoe een revisietoets op een valide en betrouwbare manier een bijdrage kan

leveren aan een grootschalige toetsing van schrijfvaardigheid in het primair onderwijs. Hiertoe wordt ten eerste de validiteit van een bestaande, gestandaardiseerde meerkeuzetoets voor revisievaardigheid geëvalueerd. Daarnaast is een open variant van deze toets ontwikkeld, waarin leerlingen hun verbeteringen in de tekst doorvoeren. In een experimentele studie zijn de validiteit en de toetskarakteristieken van de open variant van de toets vergeleken met de bestaande meerkeuzetoets.

Een revisietoets in open vraagvorm heeft als voordeel dat de toets een authentieke representatie biedt van het actief verbeteren van een tekst. Uit een analyse van de toetskarakteristieken blijkt bovendien dat de betrouwbaarheid en het discriminerend vermogen van de open variant van de toets voldoen. De domeinrepresentatie van een (papieren) open revisietoets is echter beperkt, omdat niet alle elementen uit het revisieproces geschikt zijn om objectief te toetsen in een open vraagvorm. Een meerkeuzetoets daarentegen wordt weliswaar gezien als een minder authentieke representatie vanwege de passieve bevraging, maar biedt in potentie een bredere domeindekking doordat meer elementen uit de tekst specifiek bevraagd kunnen worden (bijv. de volgorde van alinea's). De inhoudsanalyse van de opgaven in beide onderzochte toetsversies laat echter een duidelijke nadruk op formuleervaardigheid zien. De revisietoetsen zijn daarmee in hun huidige vorm geen valide representatie van het *gehele* construct revisievaardigheid, maar toetsen in plaats daarvan een enkel element van het reviseren: namelijk de mate waarin leerlingen de oppervlaktekenmerken (bijv. de juistheid van het taalgebruik) van de tekst kunnen verbeteren. Deze resultaten suggereren het gebruik van een toetsvorm waarin zowel meerkeuzevragen als open vragen gecombineerd worden om een bredere domeinrepresentatie te bewerkstelligen en daarmee bij te dragen aan een valide toetsing van revisievaardigheid.

In **Hoofdstuk 4** is de inzet van een geautomatiseerde beoordeling van opstellen (*automated essay evaluation* [AEE]) verkend. Het doel van deze exploratieve studie is het onderzoeken van de relatie tussen maten van tekstcomplexiteit en tekstkwaliteit enerzijds, en de mogelijkheden om deze maten in te zetten binnen een schrijfvaardigheidsmeting in het primair onderwijs anderzijds. Hiervoor is gebruik gemaakt van T-Scan, een programma voor geautomatiseerde complexiteitsanalyse van Nederlandse teksten (Kraf, Van der Sloot, Pander Maat, Van den Bosch & Kleijn, 2013). De resultaten laten allereerst zien dat de validiteit van verschillende maten van tekstcomplexiteit beïnvloed wordt door het feit dat schrijfproducten in het primair onderwijs van gebrekkige kwaliteit zijn—zoals geïllustreerd wordt door de opstellen in Figuur 1. Deze uitkomst benadrukt de noodzaak om een voorbewerkingsmodule (Pre-Scan) te ontwikkelen waarin fouten in grammatica en interpunctie worden gedetecteerd en gecorrigeerd.




*Figuur 1*. Voorbeelden van gebrekkig taalgebruik in schrijfproducten.

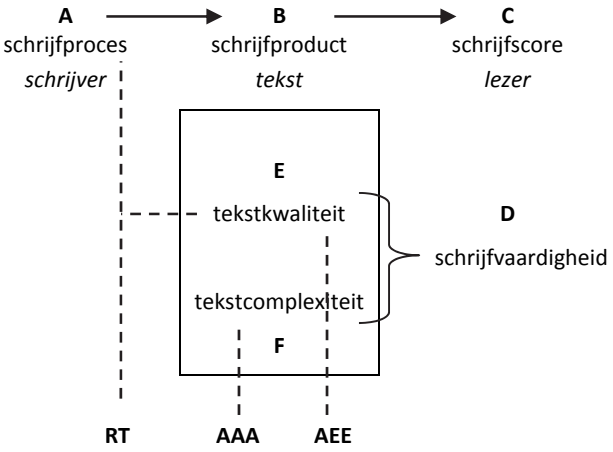
Om te bepalen welke kenmerken van tekstcomplexiteit indicatief zijn voor de schrijf-vaardigheid van basisschoolleerlingen, is een kwantitatieve analyse uitgevoerd van de overeenstemming tussen 37 maten van tekstkwaliteit uit T-Scan en twee maten van schrijfvaardigheid, te weten leerjaar en opstelscore. Op basis van de uitkomsten zijn 13 maten van tekstcomplexiteit geselecteerd. Om de interpreteerbaarheid en validiteit van deze 13 tekstkenmerken als maten van schrijfvaardigheid te evalueren is vervolgens een kwalitatieve analyse uitgevoerd. De resultaten hiervan tonen ten eerste aan dat verscheidene maten aangepast dienen te worden om ze specifiek geschikt te maken voor de evaluatie van de schrijfvaardigheid van beginnende schrijvers. Daarnaast maken de resultaten duidelijk dat verschillende factoren (zoals tekstlengte) de gemeten complexiteit van een tekst onrechtmatig beïnvloeden.

Al met al laten de resultaten van deze studie zien dat AEE kan dienen als een consistente meetmethode, waarmee een uitgebreide analyse van verscheidene linguïstische tekst-kenmerken die gerelateerd zijn aan tekstkwaliteit mogelijk wordt. De resultaten van een kwalitatieve studie benadrukken daarnaast de noodzaak om nader onderzoek uit te voeren en zo de bruikbaarheid van AEE in het primair onderwijs verder te onderbouwen. Desalniettemin kan AEE in de huidige vorm al ingezet worden om verschillen tussen groepen leerlingen te karakteriseren binnen een grootschalig onderzoek. Op deze manier biedt de evaluatie van tekstcomplexiteit een objectieve en informatieve methode om de toetsing van schrijfvaardigheid mee te verrijken en het schrijfonderwijs mee te ondersteunen.

**Hoofdstuk 5** tot slot, bevat een algehele discussie van de resultaten uit de voorgaande hoofstukken, en biedt aanbevelingen voor het toetsen van schrijfvaardigheid binnen een grootschalig peilingsonderzoek in het primair onderwijs. De resultaten uit deze studie impliceren dat de toegepaste evaluatiemethoden verschillende delen van het produceren van een tekst (het schrijf*proces*) dekken, door verschillende elementen van het resultaat van dit proces (het schrijf*product*) te evalueren. Bovendien combineert deze studie twee benaderingen van het beoordelen van schrijfvaardigheid door zowel de *kwaliteit* als de *complexiteit* van geschreven teksten in beschouwing te nemen, waardoor informatie verkregen wordt over de relatie tussen specifieke tekstkenmerken enerzijds en de vaardigheid van de schrijver anderzijds.

**Conclusie**

Deze studie had tot doel om de validiteit en betrouwbaarheid van schrijfvaardigheids-toetsing in het primair onderwijs te verbeteren. Hiertoe zijn verschillende methoden geëvalueerd om zowel de *kwaliteit* als de *complexiteit* van teksten geschreven door beginnende schrijvers te beoordelen. Figuur 2 geeft een schematische representatie van schrijfvaardigheidstoetsing en de rollen van tekstkwaliteit en tekstcomplexiteit daarbinnen. Het resultaat van het schrijfproces (A), is een schrijfproduct (B), waaraan tijdens de beoordeling een score (C) wordt toegekend. Het schrijfproduct wordt beschouwd als een uiting van schrijfvaardigheid (D) en wordt beoordeeld door zowel tekstkwaliteit (E) als tekstcomplexiteit (F) te evalueren.




*Figuur 2.* De schematische relaties tussen tekstkwaliteit, tekstcomplexiteit en schrijfvaardigheid.

In deze studie zijn drie beoordelingsmethoden besproken die elk een verschillend aspect van het construct schrijfvaardigheid evalueren, zoals geïllustreerd wordt in Figuur 2. Een revisietoets (RT, Hoofdstuk 3), toetst de mate waarin schrijvers de kwaliteit van een tekst kunnen evalueren en reviseren tijdens het schrijfproces. Een analytische beoordeling met ankeropstellen (AAA, Hoofdstuk 2) evalueert de kwaliteit van een tekst met het oog op het schrijfdoel. Binnen een automatische evaluatie van schrijfproducten (AEE, Hoofdstuk 4) tot slot, worden linguïstische complexiteitskenmerken van teksten geëvalueerd.

De resultaten zoals gepresenteerd in dit proefschrift suggereren gunstige effecten op zowel de betrouwbaarheid als de validiteit van een schrijfvaardigheidsmeting wanneer de hierboven beschreven methoden worden toegepast. Ten eerste blijkt de toevoeging van ankeropstellen aan een analytische beoordelingsprocedure een gunstig effect te hebben op de overeenstemming tussen beoordelaars, de validiteit van de schrijfscores en de mate waarin deze schrijfscores generaliseerbaar zijn over verschillende taken en beoordelaars. Daarnaast biedt een revisietoets in open vraagvorm een valide methode om te meten in hoeverre schrijvers in staat zijn de kwaliteit van een tekst te evalueren en te verbeteren. Tot

slot is de geautomatiseerde evaluatie van tekstcomplexiteit een veelbelovende objectieve methode om tekstkenmerken te evalueren die indicatief zijn voor de schrijfvaardigheid van beginnende schrijvers. Al met al blijken de onderzochte beoordelingsmethoden zowel de constructrepresentatie als de consistentie van de schrijfmeting te verhogen, en dragen de methoden daarmee bij aan een valide en betrouwbare beoordeling van schrijfvaardigheid in het primair onderwijs.

**Referenties**


Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bergh, H. van den, De Maeyer, S., van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. van Steendam, M. Tillema, G. C. W. Rijlaarsdam, & H. van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices.* Leiden/Boston: Brill.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research*, *1*(1), 1-26.

Kraf, R., van der Sloot, K., Pander Maat, H., van den Bosch, A., & Kleijn, S. (2013). *T-Scan*. Retrieved from http://languagelink.let.uu.nl/tscan

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement*. (3rd ed., pp. 13–104). Washington, DC: American Council on Education.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modelling. *Language Testing, 22*, 1-30.

## Curriculum Vitae

Hiske Feenstra was born on August 24th, 1981, in Groningen, the Netherlands. She obtained her gymnasium diploma at Lorentz College in Arnhem in 1999, and started to study Romance languages and cultures at Groningen University in 2000. After her first year, she enrolled in the post-propaedeutic programme General Linguistics. In 2005, she graduated cum laude, obtaining a master's degree in Theoretical Linguistics. During an internship at the university's expertise center for language and communication, she became acquainted with educational publishing and language assessment. After graduating, she worked as a freelance author in both fields, before starting her career in educational assessment at Cito in 2007. There, she developed a range of testing materials, and in 2009 started her PhD project on writing assessment. Besides the execution of her PhD project, she worked as an international consultant and as a national project manager for the 2012 Programme for International Student Assessment (PISA). She currently manages the development of a nationwide adaptive test in primary education at Cito, and was elected as a member of Cito's Works Council in spring 2014. In addition, she is a member of the expert panel concerned with the examination of Dutch at the Schola Europaea.