Writing in first and second language

Empirical studies on text quality and writing processes

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006

e-mail: lot@uu.nl
http://www.lotschool.nl

Cover illustration by Wouter Kortman

ISBN: 978-94-6093-082-9
NUR 616

Writing in first and second language
Empirical studies on text quality and writing processes

Schrijven in eerste en tweede taal
Empirisch onderzoek naar tekstkwaliteit en schrijfprocessen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op dinsdag 19 juni 2012 des ochtends om 10.30 uur

door

Marion Tillema

geboren op 31 maart 1981 te Groningen

Promotoren: Prof. dr. H. van den Bergh
Prof. dr. G.C.W. Rijlaarsdam
Prof. dr. T.J.M. Sanders

TABLE OF CONTENTS

ACKNOWLEDGMENTS

Carrying out this research project, and writing this dissertation, have been a wonderful experience. It has been a privilege to work with so many people who were often inspiring teachers and kind friends at the same time. What follows below is an attempt to put my gratitude towards all the people who helped me into words. Although this dissertation would not have existed without the help of all of them, any remaining errors are, of course, my own.

I thank my supervisors: Huub van den Bergh, Gert Rijlaarsdam, and Ted Sanders. Huub, thank you for teaching me so much, for being a terrific supervisor and mentor, for your patience during my detours (both in the "real world" and while writing research papers), and for putting things into perspective. I will miss our chats in the Trans courtyard, which usually involved you having a luminous idea before you could finish your cigarette. Gert, thank you for your practical and moral support. I have really enjoyed our (email) conversations about interpretations of new results. In addition, I really appreciate it that, despite the geographical distance between us, you involved me in your projects. I have learned a lot from them. And of course, this dissertation could not have developed into what it is without your comprehensive and amazingly quick feedback. Ted, I have always greatly valued your take on intermediary versions of my research papers. I have no doubt that I have greatly benefited from your feedback. In addition, I thank you for getting me involved in teaching. It has been such fun to teach *Lees- en Schrijfprocessen* (the first course you got me involved in), and other courses, too, later.

Second, I thank all my colleagues at the *afdeling Taalbeheersing*. A special 'thank you' goes to my roommates and fellow *AiO*s, without whose company and moral and practical support (remember those piles of essays?) my motivation would probably have faltered long ago. Daphne, Nina, Naomi, Ingrid, Anita, Hanna, Sanne, Rogier, Gerdineke, Anneloes, Rosie, Hans Rutger: thank you! Naomi and Hanna deserve extra-extra thanks, for agreeing to be my paranimfs.

I wish to thank UiL-OTS, and Martin Everaert and Maaike Schoorlemmer in particular, for giving me the opportunity to carry out my research in Utrecht, for allowing me to do this with a not-so-customary timeline, and for their practical support.

A large number of people helped me out during the process of collecting and processing data. My thanks go to: Daphne van Weijen, with whom I constructed many of the research materials in this project; Fleur Zijp, teacher at the *Utrechts Stedelijk Gymnasium*, and the students in forms 3a, 3b and 3d in 2008/2009, for their participation in the main experiment of this research; Hans de Zwart,

teacher at the *Christelijk Lyceum Veenendaal*, for allowing me to carry out pretests there; Martine Braaksma and Talita Groenendijk at the Graduate School of Child Development and Education of the University of Amsterdam for their technical support; Mattis, Lieke and Lisanne for being "my" research assistants; Ton Koet, David Geneste, and Eveline Karelse for their participation in the essay rating project; and finally Marielle Leijten and Luuk Van Waes for Inputlog.

Furthermore, my thanks go to the EARLI SIG Writing community for their feedback during conferences. Special thanks go to Elke Van Steendam, Martine Braaksma, and Veerle Baaijen for being such good company during conferences: thanks for all the good fun! In addition, I have really enjoyed all our writing-related conversations, and the work we have done together, or still do together.

Finally, of course, I thank my dear family. *Mama: bedankt voor het belang dat jij en papa altijd hebben gehecht aan mijn opleiding. Het was mooi geweest als papa mijn promotie had kunnen meemaken.* My sisters Frieda, Eeke, and Anneke: I am glad that I have you three. Thank you for your moral and practical support (babysitting!). And then, of course, I thank my dear husband Jeroen and our lovely son Wouter, for being my anchors. Jeroen: thanks for supporting me. Your pragmatism and flexibility (!) have helped me so much during the process of completing this project. *Wouter: je bent een lieve schat! En ik ben heel blij met jouw mooie tekening op de voorkant van mijn boek. Dankjulliewel, liefies!*

Marion

Tilburg, April 3rd 2012

CHAPTER 1

INTRODUCTION

This thesis is about writing proficiency among students of secondary education. Writing is one of the most important skills for educational success, but also one of the most complex skills to be mastered. Flower and Hayes (1980) and Hayes and Flower (1980) put the complexity of writing on the map by positing their well-known cognitive model of writing processes. This model consisted of three main components. One component is the *writer's long-term memory*, which stores the writer's knowledge, for example: topic knowledge, audience knowledge, and writing plans. Another component is the *task environment*, which includes the specifics of the assignment (intended topic and audience) and the text produced so far. Both the writer's long-term memory and the task environment, according to the model, influence the third component: the *writing process*. Four main cognitive activities are identified within the writing process component, namely planning, translating (i.e. putting ideas into language) and reviewing/revising, all three of which are regulated by a fourth cognitive activity: monitoring. The complexity of writing arises from the fact the all these components need to be attended to during writing, often simultaneously. Flower and Hayes (1980) therefore referred to writing as 'juggling with constraints'. Hayes (1996) presented an updated version of the 1980 model. The 1996 model consists of two main components: the *task environment* and the *individual*. The task environment component includes, besides the physical environment (text so far and composing medium), the social environment, which signifies that writing is a communicative act in which the writer is interacting with his audience, for instance. The individual component, like in the 1980 model, includes the writer's long-term memory: topic knowledge, audience knowledge, genre knowledge, linguistic knowledge, and task schemas (procedural knowledge about, for instance, writing strategies). In addition, the individual component in the 1996 model includes the writing process (which was a separate component in the 1980 model), and two new subcomponents, namely working memory and motivation/affect.

The models of writing forwarded by Flower and Hayes (1980), Hayes and Flower (1980) and Hayes (1996) describe the constituent parts of writing, but make no claims about, for example, which knowledge from long-term memory is, or should be, used during the writing process, or how writing processes should be organized. As such, the models do not describe the different characteristics of ‘good writing’ and ‘poor writing’ (although it may be inferred that failing to attend to any of the model’s components will result in poorer writing).

Many researchers have focused on the execution of writing processes in relation to the quality of writing. Over the last two decades, researchers have increasingly acknowledged that the quality of writing processes is reflected by *the moment at which* cognitive activities (such as planning, formulating, structuring, and revising) are applied during the writing process (Breetvelt, Van den Bergh & Rijlaarsdam, 1994; Leijten & Van Waes, 2006; Olive, Kellogg & Piolat, 2008; Roca de Larios, Marín & Murphy, 2011; Rijlaarsdam & Van den Bergh, 1996; Rijlaarsdam et al., 2011; Van den Bergh & Rijlaarsdam, 1996; Van den Bergh et al., 2009; Van der Hoeven, 1997; Van Weijen, Van den Bergh, Rijlaarsdam & Sanders, 2008). The onset of this viewpoint was probably the introduction of a ‘probabilistic model of writing processes’ by Rijlaarsdam and Van den Bergh (1996) and Van den Bergh and Rijlaarsdam (1996). They focused on the idea that the function of a cognitive activity depends on the context in which it occurs. For example: reading the assignment at the start of task execution probably serves a different function than reading the assignment during final stages of the writing process. At the start of task execution, reading the assignment is a means of determining the intended text’s topic, goals, audience, et cetera. Towards the end of task execution, reading the assignment is more probably an aspect of evaluative activities: does the produced text satisfy the assignment’s requirements? They supported these claims empirically by demonstrating that the effectiveness of cognitive activities is different during different moments of the writing process. For instance, structuring activities were more effective when they occurred during early stages of task execution (i.e. the correlation between structuring and text quality is at its highest at the start of the writing process) and less effective when they occurred towards the end of task execution (Rijlaarsdam & Van den Bergh, 1996). The researchers therefore advocated a temporal analysis of activities over the writing process: investigations of writing processes should take into account the moments at which cognitive activities occur.

Writing in L1 and L2

Due to globalization, the ability to express oneself in a language other than the first language (L1) is increasingly becoming a condition for educational success. In the Netherlands, students in secondary education learn a number of foreign languages. They are English, German, and French. Schools may sometimes substitute French or German by Spanish, Russian, Italian, Arabic or Turkish. During the final years of secondary education, when students specialize in some subjects, and drop others, Dutch and English are still mandatory school subjects. In short, English is probably the most important foreign language taught in schools in the Netherlands. As English is frequently used in Dutch media, children in the Netherlands are usually confronted with English from a young age. English is therefore not only learned in schools, but also acquired in more natural settings. It might therefore be argued that English can, in the Netherlands, probably be considered a second language (L2). In the remainder of this thesis, we will refer to it as such.

Although secondary school students are already quite able to express themselves in English (L2), their English proficiency is generally at a substantially lower level than their Dutch (L1) proficiency. Such language proficiency differences between L1 and L2 are generally thought to be the cause of the often observed quality difference between L1 and L2 writing (Sasaki & Hirose, 1996; Schoonen et al., 2003). Language difficulties are assumed to affect the quality of writing in two ways. First, students' lower L2 proficiency limits their ability to express their ideas. Second, language difficulties are believed to constrict working memory resources, leaving fewer resources for conceptual and regulatory activities (such as structuring and monitoring) and/or causing an inability to transfer L1 writing strategies to L2 writing situations. As a result, L2 texts are often of lower quality than L1 texts, in terms of language use, but also in terms of organization.

Figure 1 illustrates this problem with an L2 and an L1 essay produced by the same fifteen-year-old student of secondary education, who was a participant in the present study. Beside severe language problems, the L2 essay shows organizational errors. The writer does not choose a clear viewpoint. Almost all of the last paragraph, for example, is used for discussing disadvantages, but then the last sentence of this paragraph (and of the essay), which should be a concluding sentence, expresses that camera surveillance could be advantageous. This conclusion comes 'out of the blue'. The L1 essay shows fewer problems. Although the L1 text contains some stylistic problems, its content is fairly well organized. Pros and cons are mostly discussed in separate sections, so that the writer's line of reasoning is easily followed. In addition, the author expresses a clear position on the issue in the L1 essay.

Figure 1. Example of an L2 and an L1 text

L2 essay:

Do surveillance cameras in inner city areas increase public security?

A few time ago some people decided to bring up surveillance cameras. It shall be for the safety of people in the city or maybe the whole country. They tried it with some cities but not all of the cities agree with the good working of the surveillance cameras, for example Sneek. They said it is not going to work, so they don't want the cameras in their city.

But also a problem of the surveillance cameras is that people don't want to be watched over them, so they can stop the process. But on the other way, people can be happy that cameras hanging on the corner to look after them. For example: You live in Utrecht, in a bad neighberhoud and you walk around to the nearest supermarkt. You walk on the street and there are some guys, 20 years old, and they want to have your bag. You run away, but they are with 4 so you don't have a change. They stole your bag and they are gone. When there are cameras you can go to the police and said this and this happens, you can check it on the tapes of the cameras. Then they recognise the boys and they know how they look, etc.

Cameras don't increase public security if they are not using badly. The cameras can be used to look for a readon to hate somebody or arrest someone. That is not good if you're only checking that person. You have to watch all of them. And the cameras must have not more than 2 MegaPixels. Then they can recognise people, but not very well if they are with something they shamed for. Cameras are helping, but it is a difficult subject.

Figure 1 (continued). Example of an L2 and an L1 text

L1 essay:

<u>De mobiele telefoon: irritant of onmisbaar?</u>

Er zijn veel mensen die zich ergeren aan het gebruik van mobiele telefoons. Maar wat moet je zonder een mobiel? Sommige dingen zijn overbodig zoals smsjes: "Hi, hoe gaat het? Beetje lekker gesport? Ik ben nu aan het computeren. X." Dat zijn van die onzin smsjes, want wat kan de persoon ermee als die weet dat jij aan het computeren bent? Niet veel dus.

Maar er zijn natuurlijk ook onmisbare momenten. Stel je bent een zelfstandig ondernemer en je reist van Amsterdam naar Twello, van Rotterdam naar Utrecht iedere dag weer. Dan is het erg handig als je een mobiel hebt om met klanten in de auto te bellen en dus ondertussen te werken! Natuurlijk zijn er ook andere momenten, bijvoorbeel in de supermarkt. Volgens <u>www.bellen.com</u> belt ruim 50% van de bellers in de supermarkt. 75% van deze gesprekken gaat over het boodschappen doen. In dit soort gevallen is het handig om een telefoon te hebben om bijvoorbeeld even te vragen of je nog iets mee moet nemen. Ook is een mobiel handig om te zeggen dat je iets later komt of dat je niet komt vanwege een of andere reden.

Ik vind dat een telefoon onmisbaar is in veel gevallen. Zo ben je bereikbaar voor vervelend, maar dringend nieuws als bijvoorbeeld van de politie, maar ook voor je vriendje dat je nog iets bij de supermarkt moet kopen. Het is misschien irritant als je in de trein zit en een meisje van 14 hoort kletsen met een vriendin en alleen maar roddelverhalen verteld waar jij niet op zit te wachten. Daar heb je niet zoveel aan en dat kan storen. Maar daarvoor zijn nu stiltecoupés in de trein (volgens de inleiding van deze opdracht). Toch blijf ik erbij dat je eigenlijk altijd een mobiel mét beltegoed op zak moet hebben voor veiligheid, handigheid en gemakkelijkheid.

Figure 1 (continued). Example of an L2 and an L1 text

L1 essay translated into English (errors are translated errors from the Dutch original):

Mobile phones: irritating or essential?

There are many people who are annoyed by the use of mobile phones. But how can you survive without a mobile phone? Some things are unnecessary like text messages: "Hey, how are you? Had a good time at the gym? I'm working on the computer now. X." Those are nonsensical texts, because what use is it to someone if he knows that you are working on the computer? Not much.

But of course there are also essential moments. Imagine that you are a self-employed entrepreneur and travel from Amsterdam to Twello, from Rotterdam to Utrecht every single day. Then it is very handy if you have a mobile phone to call customers in the car and so to work while you are on the road! Of course there are also other moments, for exampl in the supermarket. According to www.bellen.com over 50% of the callers calls in the supermarket. 75% of these conversations is about the groceries. In cases like these it is handy to have a telephone to ask, for example, if you should bring anything. A mobile phone is also handy to say that you will be a bit later or that you cannot come for some reason.

I think that a mobile phone is essential in many cases. For example, you can be reached for bad, but urgent news like, for instance, from the police, but also by your boyfriend that you need to buy something at the supermarket. It might be annoying if you are on a train and overhear a 14-year-old girl chatting to a friend and gossiping, something you are not waiting to hear. That is not of much use to you and that can be intrusive. But for that reason there are now silent compartments in trains (according to the introduction of this assignment). Still, I stick to the opinion that you should really always have a mobile phone with enough call credit with you for safety, handiness and convenience.

Present study: participants and apparatus

To investigate the quality difference between L1 and L2 writing, a study was set up in which L1 and L2 writing were compared. Participants were twenty fourteen- and fifteen-year-old students of secondary education. They each wrote four short argumentative essays in L1 (Dutch), and four short argumentative essays in L2 (English). A comparison was made between L1 and L2 writing both in terms of text quality and in terms of writing processes. By using multiple tasks per language, such comparisons are warranted. After all, if only one task were used per language, it would be impossible to determine if any differences which are found are due to task or due to language (Van den Bergh et al., 2009; Van Weijen, 2009).

The quality of each text was rated by three expert raters (independently of each other). Writing processes were analyzed in terms of the following cognitive activities: reading the assignment, (process and content) planning, formulating, reading own text, evaluating own text, and revising. Students' writing processes were registered by means of think aloud procedures, combined with keystroke logging (Leijten & Van Waes, 2006). Keystroke logging has the advantage of reliable measurements: its registrations occur automatically and in a manner which does not intrude the writing process. However, keystroke logging can only provide information about cognitive activities such as formulating (typing) and revising. To obtain information about cognitive activities of a more conceptual nature, think aloud techniques can provide useful information. For instance, if a student is not typing: is he planning, is he evaluating, or is he thinking about something entirely unconnected to the writing task? By combining keystroke logging and think aloud procedures, the obtained writing process data are expected to be as reliable and complete as possible.

The execution of writing processes (and the relation between writing process and text quality) is assumed to be affected by learner variables (cf. Hayes, 1996; Van den Bergh et al., 2009). Two such learner characteristics were measured with offline tests, namely language proficiency in Dutch and English, and writing style.

The study reported in this thesis, then, looks into various constituent parts of the individual component in Hayes' (1996) model, namely cognitive activities during writing processes and knowledge in long-term memory. The two knowledge factors incorporated in this study are metacognitive knowledge of personal writing style, which probably fits into Hayes' (1996) task schema, and knowledge of language. In addition, the study relies heavily on Rijlaarsdam and Van den Bergh's (1996) and Van den Bergh and Rijlaarsdam's (1996) probabilistic model of writing, as writing processes are analyzed temporally.

Reading guide: chapters

The results of this study were reported in separate journal articles. Slightly adapted versions of these articles are presented in chapters 2 to 5. As a result, there is some overlap between the chapters, mainly between the method sections of chapters 2 to 5. The advantage of presenting four journal articles is that it is possible to read the chapters independently of each other.

The starting point of the research reported in this thesis is the quality difference between L1 and L2 texts. Although this is a widely observed phenomenon, it has not yet been possible to quantify this quality difference. Rather, researchers have sometimes compared isolated features of L1 and L2 writing. Silva (1993), for example, reports L2 texts to be shorter, to contain more linguistic errors, to contain less cohesive argumentation, and to be less focused on the reader. However, L1 and L2 text scores which express students' writing proficiency as a whole could generally not be compared. For this to be possible, the quality of both L1 and L2 texts must be expressed on the same scale. This is normally not the case, due to various causes, such as raters' different attitudes towards L1 and L2 writing, which causes them to be more strict to one of the two languages.

In **chapter 2**, a procedure is presented and tested which is expected to make direct comparisons of L1 and L2 text scores possible. The two main features of this procedure are: 1) raters are bilingual or near native users of both L1 and L2, which increases the chance that they are equally strict or lenient while rating L1 and L2 texts; 2) ratings are performed with L1 and L2 benchmark texts (i.e. texts representing average quality). The procedure is found to be successful, in that direct comparisons of L1 and L2 text scores are warranted. This chapter, then, has a methodological focus.

Since the quality of L1 and L2 texts can be expressed on the same scale, it is possible to make a direct comparison of L1 and L2 relations between writing processes and text quality. This is done in **chapter 3**, using temporal analyses. It is the aim of this chapter to find out whether effective L2 writing processes are different from effective L1 writing processes.

In **chapter 4**, language proficiency is considered in the analysis as a learner variable. The main aim of this chapter is to investigate whether language proficiency affects the manner in which the writing process is carried out, both during L1 and L2 writing. This is to be expected, particularly for L2 writing, as it has been hypothesized (e.g. Sasaki & Hirose, 1996; Schoonen et al., 2003) that language difficulties constrict working memory resources, causing the writing process to be carried out with lower quality. Conform Rijlaarsdam and Van den Bergh's (1996) and Van den Bergh and Rijlaarsdam's (1996) claim that the quality of process execution is reflected by temporal distributions of cognitive activities during writing, writing processes are analyzed temporally. Do students who are highly language proficient, for example, plan at different moments during the writing process than students who are less language proficient?

Another learner variable which possibly affects writing processes, and relations between writing processes and text quality, is writing style. Two writing styles which have been regularly described in literature on writing are the 'planner' and 'reviser' styles (Biggs, Lai, Tang & Lavelle, 1999; Kieft, Rijlaarsdam & Van den Bergh, 2006, 2008; Torrance, Thomas & Robinson, 1994, 1999, 2000). Planners are said to plan extensively before commencing text production, while revisers use text production to arrive at a text plan. Kieft et al. (2006; 2008) developed the Writing Style Questionnaire to measure the degree to which students are planners or revisers. However, as planner and reviser styles are basically characterizations of students' writing processes, it is important to investigate whether students' self reports in the Writing Style Questionnaire can predict actual writing behavior. This is done in **chapter 5**, for L1 writing. Chapter 5, then, essentially presents a methodological issue. If the Writing Style Questionnaire has sufficient predictive value, and is a valid measure of planner and reviser styles, its outcome can be used as a variable in future studies on writing processes and text quality.

Chapter 6, finally, presents a discussion of the interconnections between the findings reported in chapters 2 to 5. In addition, chapter 6 contains a section on methodological issues of the presented research, and a section with suggestions and considerations for future research.

CHAPTER 2

QUANTIFYING THE QUALITY DIFFERENCE BETWEEN L1 AND L2 ESSAYS. A RATING PROCEDURE WITH BILINGUAL RATERS AND L1 AND L2 BENCHMARK ESSAYS

Abstract [1]

It is the consensus that, as a result of the extra constraints placed on working memory, texts written in a second language (L2) are usually of lower quality than texts written in the first language (L1) by the same writer. However, no method is currently available for quantifying the quality difference between L1 and L2 texts. In the present study, we tested a rating procedure for enabling quality judgments of L1 and L2 texts on a single scale. Two main features define this procedure: 1) raters are bilingual or near native users of both the L1 and L2; 2) ratings are performed with L1 and L2 benchmark texts. Direct comparisons of observed L1 and L2 scores are only warranted if the ratings with L1 and L2 benchmarks are parallel tests and if the ratings are reliable. Results showed that both conditions are met. Effect sizes (*Cohen's d*) indicate that, while score variances are large, there is a relatively large added L2 effect: in the investigated population, L2 text scores were much lower than L1 text scores. The tested rating procedure is a promising method for cross-national comparisons of writing proficiency.

[1] This chapter is a slightly adapted version of: Tillema, M., Van den Bergh, H., Rijlaarsdam, G. & Sanders, T. (in press). Quantifying the quality difference between L1 and L2 essays. A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing*.

Several researchers have undertaken empirical studies in which they compared writing processes in the first language (L1) and a second language (L2). The general conclusion is that L2 writing processes differ from L1 writing processes. Van Weijen et al. (2008), for example, found that differences between tasks, in terms of the organization of the writing process, were larger in the L1 than in the L2. In other words, writers' writing processes varied more in L1 than in L2. Differences between L1 and L2 writing processes were also demonstrated by Chenoweth and Hayes (2001). They found that, during L2 writing, there were lower levels of fluency (i.e. number of words written per minute), decreased burst length (i.e. number of words produced every time that a new piece of text content is generated) and more instances of revision than during L1 writing processes.

While it may be expected that these differences between L1 and L2 writing processes cause differences in the quality of the output, this effect has (to our knowledge) not yet been confirmed. A number of researchers have indicated that writing in an L2 is more demanding than writing in an L1 (Roca de Larios, Manchón, Murphy & Marín, 2008; Schoonen et al., 2003; Thorson, 2000; Van Weijen et al., 2008). However, to investigate whether the more demanding task of writing in an L2 does indeed cause L2 texts to be of lower quality than L1 texts, it is necessary to quantify the quality difference between the resulting L1 and L2 texts. This would allow for relating (quantified) processing differences to (quantified) differences between L1 and L2 texts. For this to be possible, the quality of both L1 and L2 texts must be expressed on the same scale. However, a method for achieving a single scale and enabling direct comparisons of L1 and L2 texts is currently not available.

One of the first and most substantial studies into the comparability of expressions of the quality of writing in different languages was a cross-national study by the International Association for the Evaluation of Educational Achievement (IEA), reported in Gorman et al., (1988) and Purves (1992). The researchers attempted to use a single scale to express the quality of texts in different languages and from different regions. Texts from fourteen countries, written in the local L1, were rated by means of a scoring scheme which included only rating criteria which are common to all languages involved. Criteria which were assumed to hold cross-linguistically and cross-nationally were content, organization, style and tone, and overall impression. These criteria were assumed to reflect the construct of writing proficiency throughout languages and cultures, thus allowing for comparison and quantification across languages. The ratings of the fourteen essay sets (each from another country) were performed within the countries where they were written, by L1 users of the language in question. Ratings were carried out using three benchmark essays per language and per criterion. These three benchmark essays represent different scale points, namely the mean of the scale, and high quality and low quality, respectively. The other essays were rated relative to these benchmarks. As benchmark essays were selected per participating country, different benchmark essays were used in each of the fourteen countries.

The IEA researchers concluded that a comparison of text quality between languages was not possible: the rating criteria and scales were used differently by the raters from different countries. For example, whereas the raters in some countries used the entire scale, from minimum to maximum, the raters in other countries only used the upper part, from mean to maximum. While it is, strictly speaking, possible to interpret this result as indicating that all students in the latter country performed (homogeneously) well, it seems more likely that scale shrinkage occurred: the raters in the latter country did not assign any low scores, possibly due to cultural conventions. This renders the scales incomparable across countries and languages.

Van Weijen (2009) also tested a procedure designed to directly compare L1 and L2 essays. To control for problems due to different (groups of) raters (as occurred in the IEA study, cf. Gorman et al. , 1988; Purves, 1992), Van Weijen (2009) employed a rating procedure in which the raters were presented with an essay set consisting solely of essays in L2. Included in this essay set were eight L1 essays (of average quality, each on a different topic) which had previously been translated into L2. The raters were unaware of this, so that they would apply their (implicit) L2 standards to this translated L1 essay. The idea behind this procedure was that the scores assigned to the translated versions of the L1 in the L2 rating session would probably be higher than their original scores (as assigned during the L1 rating session). This (mean) score difference would be an indication of the difference between the L1 and the L2 rating scale, thus providing the researchers with a number with which to transform the L2 scale onto the L1 scale, or vice versa. However, for two out of the eight topics, the score L1-L2 difference was much larger than for the other six tasks. This could be due to an interaction effect of language and task (i.e. the quality difference between and L1 and an L2 text is different for different tasks) or due to translation inconsistencies. Van Weijen (2009) notes that the translation of essays is a difficult task, not in the least because (language) errors had to be translated too. She concludes that this procedure did not result in a scale on which to place both the L1 and L2 essays and compare them directly.

To neutralize all these problems, Van Weijen (2009, p. 171) suggests ratings by raters who are highly proficient speakers of both languages, preferably bilinguals, a procedure which is also suggested in Purves, Gorman and Takala (1988, p. 51). This might tackle two rater problems. First of all, it is expected that bilingual raters are equally strict towards both languages. Raters who, unlike bilinguals, are more proficient in one of the two languages than in the other might not be equally strict towards both languages, i.e. they might consistently assign higher scores to one of the languages (cf. Van den Bergh & Klein Gunnewiek, 2009, who found that raters awarded higher scores to texts if they were L1 users of the language of the text than if they were L2 users of the language), even if true scores are equal across languages. This bias is expected to be less prominent, or absent, in bilingual raters. In addition, a rater who is equally proficient in both languages is expected to be more likely to apply rating standards equally across languages, that is, to rate L1 and L2 texts with equal reliability, making the ratings expressible on the same scale and therefore comparable.

In the present study, then, we tested whether ratings by bilinguals of L1 and L2 essays are indeed expressible on one scale by implementing a procedure in which ratings are carried out with both L1 and L2 benchmark essays. The main aim of the present study is to investigate whether the allocation of scores to L1 and L2 essays is similar with L1 and L2 benchmarks. If so, this would mean that there is no evidence to assume that the raters rated differently for different languages. The aim of the present study would then be warranted: to quantify the quality differences between L1 and L2 writing.

METHOD

Participants and procedures

Obtaining essays

One hundred and sixty short essays (about 250 to 300 words) were rated and compared in the present study. In line with the IEA study of written composition (Gorman et al., 1988; Purves, 1992), the essays were written by fourteen- and fifteen-year-old students (N = 20; 10 female and 10 male). In addition, fifteen-year-olds are the target population of the international PISA assessments. Should a PISA assessment of writing literacy be set up, then tools for cross-national comparisons of writing become relevant, too. The participants were from three different third-year-forms at the same school for pre-university secondary education. They were recruited by means of a call for volunteers, which was distributed by their Dutch teacher. All participants were native speakers of Dutch. They received a small financial reward for their participation. Parental consent was obtained.

Each student wrote four short argumentative essays in Dutch (L1) and four essays in English (L2). Multiple tasks were used per language, in order to be able to disentangle task effects and language effects. After all, if only one task were used per language, it would be impossible to know if any differences which are found are due to task or due to language (Van den Bergh et al., 2009; Van Weijen, 2009). All eight writing assignments were similar in terms of audience (peers), medium (a school-related magazine for secondary school students) and purpose (to convince the readers of your point of view), and differed only in terms of topic. An example of an assignment can be found in Appendix A.

The available time for each essay was approximately thirty minutes, although participants were allowed to go on longer if they felt that their essays were not finished yet. No participant used more than forty minutes. The students completed the essays during two separate days. Between the four tasks completed per day, participants were given a short break of about ten to fifteen minutes.

To avoid sequence effects and to control for effects of topic, several measures were taken. It has been established that topic can greatly influence the quality of writing (Godshalk, Swineford & Coffman, 1966; Schoonen, 2005; Van Weijen, 2009). Therefore, to disentangle language effects from topic effects, the topics were systematically balanced across languages, so that each topic occurred in both L1 and L2. So, writer 1 wrote on topics 1, 2, 3 and 4 in L1 and on topics 5, 6, 7 and 8 in L2, whereas writer 2 wrote on topics 2, 3, 4 and 5 in L1 and on topics 1, 6, 7, and 8 in L2, and so forth. The order in which L1 and L2 essays were written was also balanced across participants: ten students first completed four L1 essays, and then four L2 essays; the other ten students first completed four L2 essays, and then four L1 essays.

The essay set was randomly divided into eight subsamples. Each subsample contained twenty essays. As the complete essay set contained L1 (Dutch) and L2 (English) essays, so could - and did - each subsample. In addition, each subsample contained essays on various topics. After all, eight topics were used in the present study.

Benchmark essays and scoring guide

Six benchmark essays were selected, one for each criterion per language. The positive effect of using benchmark essays on scale reliability was advocated and demonstrated by Blok (1985), Kuhlemeier and Van den Bergh (1988), Purves (1992) and Schoonen (2005). The benchmark essays were selected from an essay set from a previous study (data collected by Rijlaarsdam & Van den Bergh, 1996, University of Amsterdam, cf. Couzijn, Van den Bergh & Rijlaarsdam, 2002). The essays in this 1996 set were written on the basis of the same assignments as used in the present study, and by students of the same age as the participants in the current study. The benchmark essays represented the (approximate) average essay quality for the rating criterion in question. They were selected by two experienced raters, after elaborate inspection of both the 1996 essay set and the current essay set. All benchmark essays were essays on which the two raters were in agreement that they were of average quality. That is, they had to represent average quality for the specific rating criterion. In addition, they should not represent any extremes for the other two criteria. For example, essays which were average in terms of structure, but very poor or very good in terms of language use, could not be used as benchmarks for the 'structure' criterion (and neither for the language criterion). If benchmarks are of approximate average quality, this enhances the reliability of ratings. After all, if a benchmark essay represents an extreme on the scale (e.g. high quality), it becomes harder for the raters to rate the essays at the other extreme (e.g. low quality) reliably. As such, the suitability of the selected benchmarks is checked after the ratings by inspecting the reliabilities of the ratings.

Ratings

Eight raters, who were not among the two raters who selected the benchmark essays, were involved in this study. They were (near) native speakers of Dutch and English. They all had the Dutch nationality, but were also highly proficient in and familiar with the English language and its conventions, through years of personal and/or professional experience. All raters used English as a main language of communication during their education and work and were familiar with text conventions in both languages. They worked as teachers (one rater worked as a teacher of English at a regular school for secondary education and did a university major in English and had lived in the U.K.; two raters taught in bilingual education) or in academic settings where English is used as one of the two main languages of professional communication (next to Dutch), both productively and receptively, on a daily basis. Three of them had also spent significant parts of their lives in English speaking communities (U.K. and U.S.A., in this case). The raters were financially compensated for their work.

The essays were first rated on global quality. Global quality ratings should reflect the quality of each essay as a whole. They are, in other words, holistic ratings. Schoonen (2005) demonstrated that holistic ratings ("collected with essay scales") have higher generalizability than analytic scores ("with scoring guides"), thereby reinforcing White's (1984; 1985, as cited by Weigle, 2002) claim that holistic ratings are preferred because of their validity. Because it has been observed that second language writers often develop different aspects of writing – such as grammar, style, and argumentation – asynchronously (Weigle, 2002, p. 120), the essays were also rated on two selected criteria, namely structure and language (see Appendix B for definitions of all three rating criteria used in the present study). These two criteria are often used in assessment studies and are typically aspects on which L2 writing quality develops differently (cf. Weigle, 2002, p. 120) than L1 writing quality. It should be noted that global quality comprises, among other aspects, structure and language (cf. Bae & Bachman, 2010, who argue that “content” comprises “coherence and organization”, the latter being similar or equal to the structure criterion in the present study). It is therefore to be expected that ratings of global quality overlap with ratings of structure and language. Structure and language, on the other hand, are assumed to be separate aspects of writing quality. To minimize the possibility that the ratings for one criterion affect the ratings on the other criteria, the ratings were carried out in three different rounds (on different days, one round per rating criterion). During each new round, the ratings for previous criteria were no longer available to the raters.

In addition to a definition of each criterion, the raters were provided with an explanation of what was 'average' about the benchmark essay for the specific criterion in question, including passages from the benchmark essay (see Appendix C). This procedure served to maximize interrater reliability. The raters had to award a score to each essay which expressed how much better or worse it was than the benchmark essay (cf. Blok, 1985), which was given the randomly set score of 100. If an essay was awarded a score of 200, for example, this meant that the rater thought it was twice as good as the benchmark essay. If an essay received a score of 50, it meant that the rater thought it was half as good as the benchmark essay.

For efficiency reasons, we implemented a "design of overlapping rater teams" (Van den Bergh & Eiting, 1989, p. 1). In such a design, the raters do not rate all texts in the data set. Instead, each rater rates a randomly selected sample of texts (for example, half of all the available texts, or eighty out of a hundred available texts). By creating overlap, it is possible to estimate rater reliabilities. In the present study, each rater rated six out of the eight available subsamples. All raters worked in two conditions: in one condition they rated three subsamples (= 60 essays, on various topics, some written in L1, some written in L2) relative to an L1 benchmark essay (see table 1), while in the other condition they rated three other subsamples relative to an L2 benchmark essay. Each rater rates each essay only once (per criterion). That is, he or she never rates the same essay relative to both an L1 and an L2 benchmark. Each essay was rated by three raters relative to an L1 benchmark (on each of the three rating criteria – where the ratings of different criteria took place on different days), and relative to an L2 benchmark by three other raters.

For example, participant 1’s essay on the topic ‘camera surveillance’ belonged to subsample 3. As can be inferred from table 1, this essay was rated relative to an L1 benchmark by raters 3, 4 and 5 (who form a jury), on all three rating criteria (global, structure, language). It was rated relative to an L2 benchmark by raters 6, 7 and 8 (who form a jury), again on all three rating criteria.

Table 1. Allocation of subsamples across raters. Each subsample contained L1 and L2 essays. L1 = ratings with an L1 benchmark; L2 = ratings with an L2 benchmark

	Subsample							
	1	2	3	4	5	6	7	8
Rater 1	L1			L2	L2	L2	L1	L1
Rater 2	L2	L2			L1	L1	L1	L2
Rater 3	L1	L1	L1			L2	L2	L2
Rater 4	L2	L2	L1	L1	L1			L2
Rater 5	L1	L1	L1	L2	L2	L2		
Rater 6		L2	L2	L2	L1	L1	L1	
Rater 7		L1	L2	L1	L2		L2	L1
Rater 8	L2		L2	L1		L1	L2	L1

The order in which benchmarks were used was balanced across raters: half of the raters (raters 1, 3, 5 and 7) performed the ratings with the L1 benchmarks first, and the ratings with the L2 benchmarks second; the other half (raters 2, 4, 6 and 8) performed the ratings with the L2 benchmarks first and the ratings with the L1 benchmarks second.

Appendix D contains a procedural manual with a step-by-step description of the rating procedure described above.

Analyses

Confirmatory factor analysis was conducted using Lisrel 8.16 to examine the latent structure of the ratings in both conditions (L1 and L2 benchmarks) simultaneously. Both invariant as well as variant restrictions were placed on the parameters (see Figure 1). The three main steps in the analysis are described below.

1. Restrictions of invariance across subsamples, within benchmark languages

Restrictions of invariance are imposed on parameters within the two conditions, that is, the sets of essays rated with L1 benchmarks and the set of essays rated with L2 benchmarks. As each subsample in the essay set was randomly assembled of twenty (L1 and L2) essays and the raters did not know to which subsample an essay belonged (and were unaware that there were any subsamples at all), it follows that

Figure 1. Invariant and variant restrictions imposed within and across conditions, r = rater.

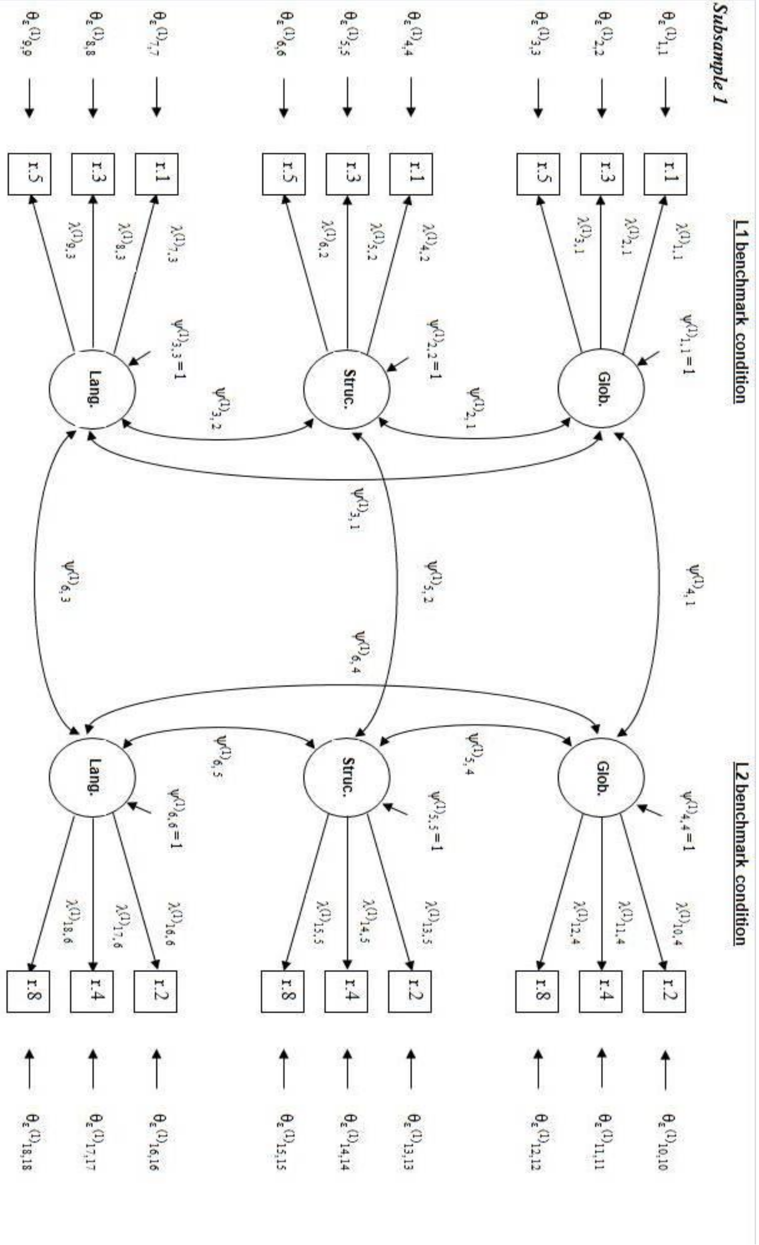


Figure 1 (continued). Invariant and variant restrictions imposed within and across conditions.

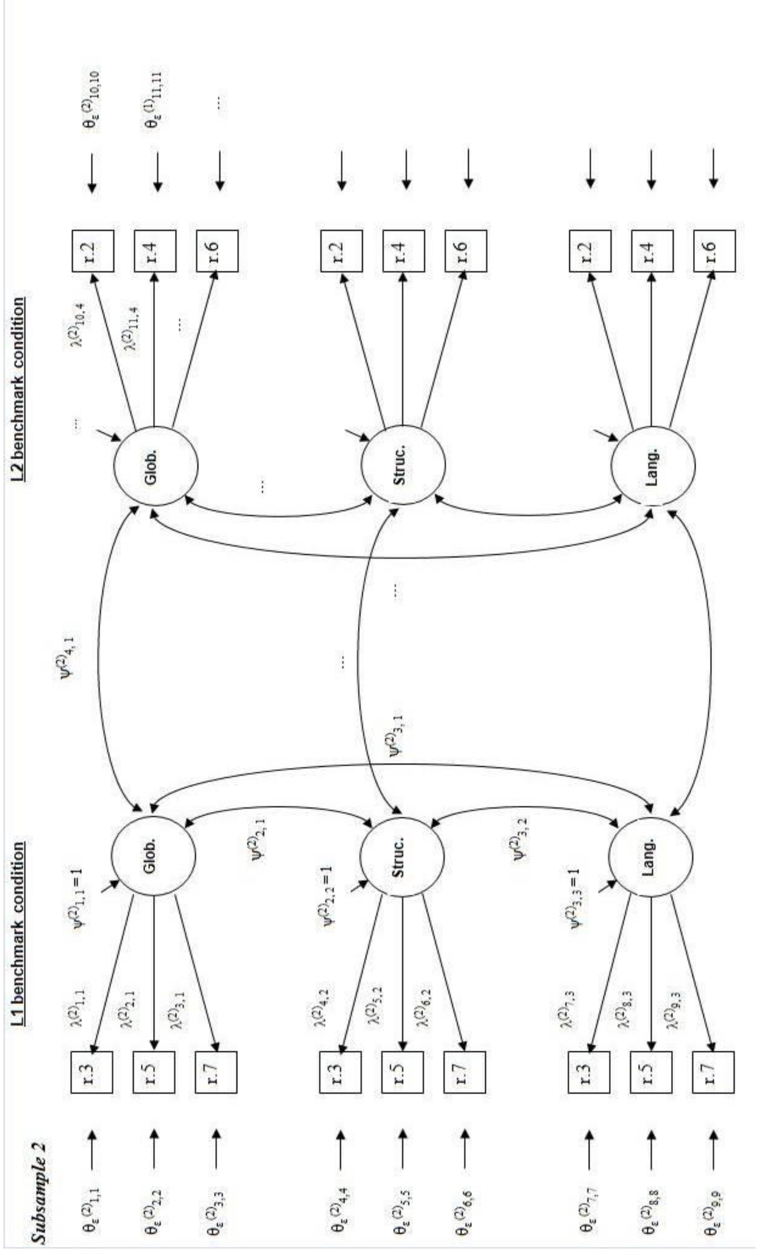
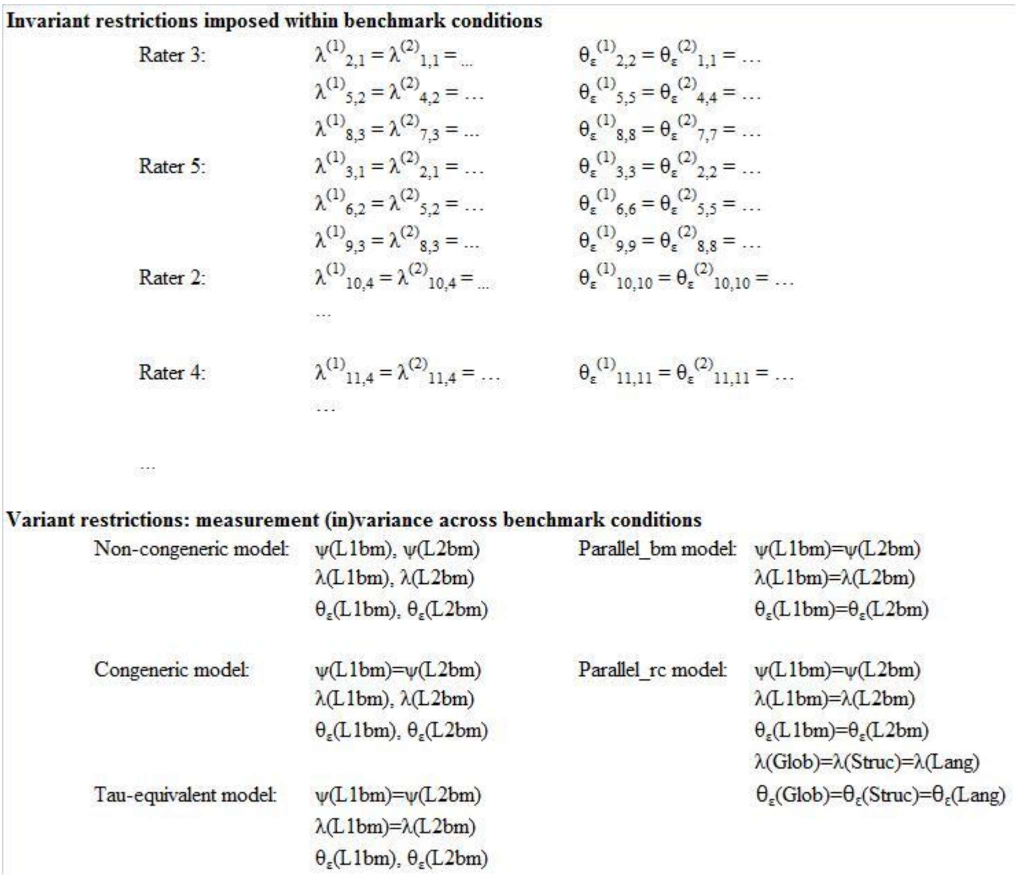


Figure 1 (continued). Invariant and variant restrictions imposed within and across conditions.



Invariant restrictions imposed within benchmark conditions

Rater 3:	$\lambda^{(1)}_{2,1} = \lambda^{(2)}_{1,1} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{2,2} = \theta_{\varepsilon}{}^{(2)}_{1,1} = \ldots$
	$\lambda^{(1)}_{5,2} = \lambda^{(2)}_{4,2} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{5,5} = \theta_{\varepsilon}{}^{(2)}_{4,4} = \ldots$
	$\lambda^{(1)}_{8,3} = \lambda^{(2)}_{7,3} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{8,8} = \theta_{\varepsilon}{}^{(2)}_{7,7} = \ldots$
Rater 5:	$\lambda^{(1)}_{3,1} = \lambda^{(2)}_{2,1} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{3,3} = \theta_{\varepsilon}{}^{(2)}_{2,2} = \ldots$
	$\lambda^{(1)}_{6,2} = \lambda^{(2)}_{5,2} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{6,6} = \theta_{\varepsilon}{}^{(2)}_{5,5} = \ldots$
	$\lambda^{(1)}_{9,3} = \lambda^{(2)}_{8,3} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{9,9} = \theta_{\varepsilon}{}^{(2)}_{8,8} = \ldots$
Rater 2:	$\lambda^{(1)}_{10,4} = \lambda^{(2)}_{10,4} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{10,10} = \theta_{\varepsilon}{}^{(2)}_{10,10} = \ldots$
	…	
Rater 4:	$\lambda^{(1)}_{11,4} = \lambda^{(2)}_{11,4} = \ldots$	$\theta_{\varepsilon}{}^{(1)}_{11,11} = \theta_{\varepsilon}{}^{(2)}_{11,11} = \ldots$
	…	
…		

Variant restrictions: measurement (in)variance across benchmark conditions

Non-congeneric model:	ψ(L1bm), ψ(L2bm) λ(L1bm), λ(L2bm) θ_{ε}(L1bm), θ_{ε}(L2bm)	Parallel_bm model:	ψ(L1bm)=ψ(L2bm) λ(L1bm)=λ(L2bm) θ_{ε}(L1bm)=θ_{ε}(L2bm)
Congeneric model:	ψ(L1bm)=ψ(L2bm) λ(L1bm), λ(L2bm) θ_{ε}(L1bm), θ_{ε}(L2bm)	Parallel_rc model:	ψ(L1bm)=ψ(L2bm) λ(L1bm)=λ(L2bm) θ_{ε}(L1bm)=θ_{ε}(L2bm) λ(Glob)=λ(Struc)=λ(Lang) θ_{ε}(Glob)=θ_{ε}(Struc)=θ_{ε}(Lang)
Tau-equivalent model:	ψ(L1bm)=ψ(L2bm) λ(L1bm)=λ(L2bm) θ_{ε}(L1bm), θ_{ε}(L2bm)		

raters should rate all subsamples in the same manner (i.e. with equal reliability) for each of the three subsamples which he or she rates per benchmark language.

Within classical test theory, differences between subsamples (and within raters) can be interpreted as random error (cf. Van den Bergh & Eiting, 1989). Therefore, invariant restrictions are imposed on ratings of the three subsamples per benchmark language performed by each individual rater. That is, the ratings of each of the three subsamples (or groups: *g*) per benchmark language by a rater (*r* = 1, 2, …, 8) have an equal regression on the true scores ($\lambda^{g}_{r} = \lambda^{g'}_{r} = \lambda^{g''}_{r}$) and an equal error variance ($\theta^{g}_{\varepsilon,r} = \theta^{g'}_{\varepsilon,r} = \theta^{g''}_{\varepsilon,r}$). These two parameters together indicate the reliability of each rater ($\varrho = \lambda^2/[\lambda^2+\theta_\varepsilon]$). These invariant restrictions hold within rating criteria.

2. Measurement (in)variance across benchmark languages

The next step is to evaluate if raters are able to carry out ratings similarly across benchmark languages. If it turns out that they are, then there is no indication that the raters rated differently for different languages. Such a result therefore allows for comparison of L1 and L2 essay scores, as obtained from the current rating procedure. To test whether ratings are indeed stable across benchmark languages, a test of measurement invariance (Jöreskog, 1971) is conducted: do raters rate the same construct, in the same way (i.e. with equal reliability), no matter the language of the benchmark essay? A χ^2 statistic is used to evaluate the absolute fit and the difference in fit of five nested models posing increasing numbers of restrictions across benchmarks.

The first model is the *non-congeneric* model. This model allows for measurement variance: the correlation between scores on global quality, structure and language is different if the benchmark essay has a different language. If this model fits the data, it means that using a benchmark essay in a different language affects rater's conceptions of the construct to be assessed. The next four models are all models of measurement invariance.

In the *congeneric* model, using a benchmark essay in a different language does not influence rater's conceptions of the construct to be assessed. Correlations between true scores on global quality, structure and language are invariant across benchmark languages, i.e. the correlations between ratings of global quality (*gq*), structure (*s*) and language (*l*) are equal in the L1 benchmark and L2 benchmark condition. Note that this implies restrictions on the correlations between different rating criteria across benchmark languages, for instance $\psi^{g}{}_{gq(L1),s(L2)} = \psi^{g}{}_{s(L1),gq(L2)}$. The congeneric model does allow for differences in regressions on true scores (λ) and error score variances (θ_ε). Although the reliability of each rater is equal across the three subsamples per benchmark language, this model involves two reliabilities per rater: one for essays rated with L1 benchmarks, one for essays rated with L2 benchmarks.

The *tau-equivalent* model is more restrictive. In this model, both correlations and true score variance are equal across benchmark languages. Therefore, the regression of the observed scores on each of the three rating criteria (global quality (*gq*), structure (*s*) and language (*l*)) on the true scores is equal across benchmark languages, for each rater. In other words, the regression of the rating criterion "global quality" on the true scores with L1 benchmarks is equal to the regression of the rating criterion "global quality" on the true scores with L2 benchmarks (i.e. $\lambda^{g}{}_{r}$,

$gq(L1) = \lambda^{g}_{r}, gq(L2))$. In the tau-equivalent model, error variance is not identical across benchmark languages. This could be the case, for example, if raters find rating essays which are in a different language than the benchmark essays (i.e. L1 essays with an L2 benchmark or L2 essays with an L1 benchmark) more difficult than rating essays in the same language as the benchmark.

The *parallel_bm* model is even more restrictive. In addition to equality of correlations and true score variance, it assumes equal error score variance across benchmark languages (i.e. $\theta^{g}_{r}\ (L1) = \theta^{g}_{r}\ (L2)$). Regression on the true score (λ) and error score (θ) together indicate the reliability of each rater ($\varrho = \lambda^2/[\lambda^2+\theta_\varepsilon]$). It follows, then, that in the parallel_bm model, individual rater reliabilities are invariant across benchmark languages (but may be variant across the rating criteria within benchmark languages, e.g. the reliabilities for "global quality" with L1 benchmarks may be different from the reliabilities for the "structure" criterion rated with L1 benchmarks). Note that the parallel_bm model, while disallowing varying reliabilities within raters, does allow for different reliabilities between raters.

The final *parallel_rc* model ("rc" for rating criteria) is an extended version of the parallel_bm model, which imposes yet another restriction. In this model, regressions on the true score and variance of error scores are not only invariant across benchmark languages, but also across rating criteria (i.e. $\lambda^{g}_{r,gq} = \lambda^{g}_{r,s} = \lambda^{g}_{r,l}$ and $\theta^{g}_{r,gq} = \theta^{g}_{r,s} = \theta^{g}_{r,l}$ where "g" is *global quality*, "s" is *structure*, "l" is *language*). That is, for each rater, the reliability of the rating of global quality is equal to the reliabilities of the ratings of structure and language, in either language.

Comparison of observed L1 and L2 scores as collected with L1 and L2 benchmarks is warranted under two conditions. First, the ratings with L1 and L2 benchmarks should be parallel tests (i.e. the parallel_bm or parallel_rc model is accepted). After all, only then can we assume that the scores are "represented by numbers on the same scale" (Jöreskog, 1971, p. 109). Second, the ratings must have been carried out with sufficient reliability in order for comparisons to be meaningful. Rater and jury reliabilities, as estimated within the preferred model, must therefore be inspected.

Whereas the degree to which the three rating criteria (global quality, structure and language) correlate is not directly relevant for answering the main question of the present study (i.e. the strength of these correlations is not relevant for deciding whether ratings with L1 and L2 benchmarks are parallel tests – the correlations only need to be similar, not necessarily high or low), it is nevertheless interesting to inspect these correlations. After all, the degree to which these three criteria correlate may (partly) reflect their validity. Although it is to be expected that overall quality is related to both structure and language to some degree, for it to be valid to distinguish between these three criteria, the correlations between the ratings of the three criteria should not equal 1. This applies particularly to the correlation between the ratings of structure and language.

3. Comparison of L1 and L2 essay scores
If a parallel model is found to best fit the observed data, and if the ratings are found to have been carried out with sufficient reliability, then a comparison of L1 and L2 essay scores is warranted. As the main question of the present study is to test the usability of the applied rating procedure with L1 and L2 benchmarks and raters who are (near) native speakers of both the L1 and the L2 for quantifying the quality difference between L1 and L2 essays, it is also of interest to investigate if it discriminates between L1 and L2 essays.

The size and significance level of the difference between L1 and L2 scores was established by submitting them to multilevel regression analysis, in which the language of the essay is the predictor variable. This model provides a distinction between the variance due to participant (i.e. differences between the averages of each writer), the variance due to topic (i.e. differences between topics) and residual variance (i.e. random error). The L2 scores are treated as the intercept, relative to which deviations due to a different language (i.e. essays in the L1) are modeled.

RESULTS

Measurement (in)variance across benchmark languages

Table 2 features chi-square statistics for the absolute fit of the five tested models, as well as for the comparison of the fit of the five models. The absolute fit of four of the five models is satisfactory. The non-congeneric model, the congeneric model, the tau-equivalent model and the parallel_bm model are good fits to the data: none of the models significantly deviate from the observed data (p > 0.05 for all models). Therefore, there is no reason to assume that the models do not fit the data. The parallel_rc model, however, does not fit the data (p < .001). Comparison of the five models shows that the fit of the congeneric model (which imposes the restriction that correlations between rating criteria should be equal across benchmark languages) is not significantly (p = .897) different from the fit of the non-congeneric model (which does not impose any restrictions of invariance on the correlations between rating criteria). Similarly, moving from the congeneric model to the more restrictive tau-equivalent model does not significantly change the fit (p = .997). The same applies for the parallel_bm model: though it is more restrictive than the tau-equivalent model, its fit is not significantly different from the fit of the tau-equivalent model (p = .912). The parallel_rc model, however, fits the data less well than the parallel_bm model (p = .001). As the parallel_rc model fits the data poorly, and because there is a significant difference in fit when compared to the other models, this model cannot be accepted. In other words, the assumption must be rejected that true and error score variance is equal across rating criteria. Therefore, the parallel_bm model in this case provides the most parsimonious description of the data.

Table 2. Absolute fit and comparison of nested models testing the degree of measurement invariance between L1 and L2 benchmark essays

Absolute fit of the five models	χ^2	df	p
Non-congeneric	1231.02	1152	0.052
Congeneric	1320.89	1260	0.11
Tau-equivalent	1330.31	1284	0.18
Parallel_bm	1345.60	1308	0.089
Parallel_rc	1779.10	1354	0.001

Comparison of models	χ^2	df	p
Non-congeneric vs. congeneric	89.87	108	.897
Congeneric vs. tau-equivalent	9.42	24	.997
Tau-equivalent vs. parallel_bm	15.29	24	.912
Parallel_bm vs. parallel_rc	433.50	46	.001

As the parallel_bm model is accepted, it may be assumed that the distribution of scores across L1 and L2 essays, as allocated by the raters, is not different for different languages of the benchmark essay. For all eight raters, the regression on the true score and residual variance, and thus reliabilities, are equal across benchmark languages. There is, in other words, no evidence to assume that the raters were unable to apply rating standards equally across languages. This result therefore allows for comparison of the scores of L1 and L2 essays.

Parameters estimated within the parallel_bm model: rater reliabilities and correlations between criteria

For the comparisons to be meaningful, it is essential that the ratings are carried out with sufficient reliability. All eight subsamples were rated by three raters for each benchmark language, who, in effect, form a jury. The jury number equals the number of the subsample which was rated by that particular group of raters. Hence, jury 1 for L1 benchmarks consists of different raters than jury 1 for L2 benchmarks. Table 3 shows the jury reliabilities, as estimated within the preferred parallel_bm model, for each rating criterion. In the L1 benchmark condition, the jury who rated subsample 1, for example, performed the rating for global quality with a reliability of .76. In the L2 benchmark condition, the jury who rated subsample 1 (consisting of three different raters than in the L1 benchmark condition) performed the global quality rating with a reliability of .68. Overall, thirty-six out of the forty-eight calculated jury reliabilities exceed .70. Only one reliability is lower than .60. In short, the ratings were performed with adequate reliability.

Table 3. Jury reliabilities for L1 and L2 benchmarks (b.m.), as estimated within the preferred parallel model, as well as standardized regressions on the true scores (λ; standard errors between brackets) and on individual rater reliabilities (ϱ), presented per rating criterion

L1 b.m.				L2 b.m.			
Jury	**Global**	**Structure**	**Language**	**Jury**	**Global**	**Structure**	**Language**
1	.76	.76	.81	1	.68	.65	.74
2	.80	.71	.80	2	.73	.68	.79
3	.74	.71	.85	3	.75	.51	.70
4	.71	.60	.71	4	.75	.67	.78
5	.73	.68	.79	5	.78	.72	.73
6	.73	.60	.72	6	.76	.76	.81
7	.76	.73	.74	7	.76	.63	.74
8	.74	.64	.65	8	.74	.76	.83

	Global quality		Structure		Language	
Rater	**λ (se)**	**ϱ**	**λ (se)**	**ϱ**	**λ (se)**	**ϱ**
1	.69 (.11)	.48	.76 (.11)	.58	.65 (.11)	.42
2	.74 (.11)	.55	.78 (.11)	.61	.70 (.12)	.49
3	.75 (.11)	.56	.74 (.11)	.55	.85 (.12)	.72
4	.60 (.12)	.36	.63 (.12)	.40	.79 (.12)	.62
5	.72 (.12)	.52	.63 (.11)	.40	.79 (.11)	.62
6	.72 (.11)	.52	.48 (.13)	.23	.75 (.12)	.56
7	.79 (.11)	.62	.63 (.12)	.40	.62 (.12)	.38
8	.59 (.12)	.35	.39 (.13)	.15	.59 (.13)	.35

Table 3 also features individual rater reliabilities, although they are less relevant than the jury reliabilities (cf. Gebril, 2009; Raymond, 1982; Schoonen, 2005; Van den Bergh & Eiting, 1989; Weigle, 2002, who all indicated that assessments by single raters are insufficient). λ indicates raters' standardized regressions (all of them significant: $\lambda > 2*se$) on the true scores, that is, the degree to which the observed scores, as assigned by the raters, relate to true scores[2]. ϱ reflects individual rater reliabilities. We can infer, for example, that rater 1 performed the ratings for global quality with a reliability of .48, indicating that this rater agreed only moderately with other raters.

The disattenuated[3] correlations between the three rating criteria (global quality, structure, and language) are shown in table 4. All correlations are strong, suggesting that the rating criteria were not fully distinguishable. Since global quality comprises structure and language (plus some other criteria), reasonably strong correlations between the ratings of global quality one the one hand and structure and language on the other hand were expected. Nevertheless, the perfect correlation between global quality and structure is a striking finding, which will be revisited in the Discussion.

Table 4. Disattenuated correlations between rating criteria. Standard errors between brackets

	Global quality	Structure	Language
Global quality	1		
Structure	1 (.05)	1	
Language	.89 (.06)	.75 (.08)	1

Comparison of L1 and L2 essay scores

Now that parallelism and sufficient reliability have been established, the observed L1 and L2 essay scores can be compared: the quality difference between Dutch and English essays can be quantified. The results of this analysis are presented in table 5 for each of the three rating criteria.

The positive regression weights (*b*-values) overall indicate that L1 texts have, on average, significantly higher scores than L2 texts for each criterion. When rated with L1 benchmarks, for example, L1 essays are on average awarded scores 44.08 points higher on global quality than L2 essays. For all other criteria, too, the L1 scores are (on average) higher than the L2 scores. Effect sizes (*Cohen's d)* are reported in table 6, indicating the size of the difference between L1 and L2 essay scores relative to the variance due to participant ($ES_{participant}$), variance due to topic (ES_{topic}) and total variance (ES_{total}). For example, $\mathrm{ES}_{\mathrm{total}}$ (global quality, L1 benchmark) equals $44.08/\sqrt{(898.57+62.28+1780.58)}$, equals $44.08/\sqrt{2741.43} = 44.08/52.36$, equals 0.84.The size of the L1-L2 score difference relative to the total

[2] Note that the true score can be viewed as the shared part of the ratings by all raters.

[3] That is, correlations between true scores, ridded of measurement unreliability.

variance ranges between .55 and 1.43. This indicates that, while score variances are large, there is a (relatively) large added L2 effect. It is also interesting to observe

Table 5. Regression results for the difference between L1 (Dutch) and L2 (English) essay scores, per benchmark language: intercepts (L2 scores), gradients (L1 scores; b value), standard errors (se), and variances (s^2) due to participant and topic of the assignment, as well as residual variance. $p < .05$.

Note 1. The different intercept values between criteria cannot be interpreted. The three criteria were rated on three different occasions and each criterion had its own benchmark essay. In addition, the fact that the parallel_rc model was a poor fit of the data, indicates that the ratings of the three different criteria are not parallel tests. Their scales are therefore not comparable.

Note 2. As ratings with L1 and L2 benchmarks have been established to be parallel tests, the differences between values of intercepts, gradients and variances are (if compared per rating criterium) not significantly different.

	L1 benchmark					
	Global quality		**Structure**		**Language**	
	estimate	**(se)**	**estimate**	**(se)**	**estimate**	**(se)**
Intercept (L2 essays)	102.00	8.67	94.80	7.27	79.96	4.5
b (deviation due to L1 essays)	44.08	8.69	23.36	5.49	40.42	3.61
s^2 (participants)	898.57	355.88	428.04	184.13	274.34	107.66
s^2 (topics)	62.28	79.94	131.27	67.15	0	0
s^2 (residu)	1780.58	218.67	1192.76	146.46	521.72	62.36
	L2 benchmark					
	Global quality		**Structure**		**Language**	
	estimate	**(se)**	**estimate**	**(se)**	**estimate**	**(se)**
Intercept (L2 essays)	105.49	8.97	93.30	5.79	93.86	6.94
b (deviation due to L1 essays)	40.59	6.24	18.79	4.54	55.01	6.21
s^2 (participants)	1019.59	384.20	294.74	126.46	558.30	238.80
s^2 (topics)	80.60	32.29	68.44	33.85	76.64	33.85
s^2 (residu)	1544.12	189.57	816.51	100.28	1540.54	189.32

that the L1-L2 difference adds substantially to the essay score differences between participants (*Cohen's d* > 1 in all cases) and between topics (*Cohen's d* = 5.59 (global quality, L1 benchmark), 0.73 (structure, L1 benchmark), 4.52 (global quality, L2 benchmark), 8.28 (structure, L2 benchmark), 6.28 (language, L2 benchmark)). From this, it can be inferred, for example, that the difference between L1 and L2 essay scores for global quality is the size of approximately five standard deviations of the score differences due to topic. So, although there are variations according to rating criterion and benchmark language, the added language effect on essay scores is very large overall. In short, the applied rating procedure – with L1 and L2 benchmarks and raters who are (near) native speakers of both the L1 and the L2 – discriminates between L1 and L2 essays and quantifies the quality difference between the two.

Table 6. Effect sizes (Cohen's d) indicating the substantiveness of the difference between L1 and L2 essay scores relative to the variance due to participant (ES$_{participant}$), variance due to topic (ES$_{topic}$) and total variance (ES$_{total}$). Effects sizes are presented per rating criterion and benchmark language
Note: As no variance was established between scores on the "language" criterion with L1 benchmarks (see table 5) due to topic, no effect size is presented for this condition in table 6.

	L1 benchmark			**L2 benchmark**		
	Global quality	**Structure**	**Language**	**Global quality**	**Structure**	**Language**
ES$_{participant}$	1.47	1.13	2.44	1.27	1.09	2.33
ES$_{topic}$	5.59	0.73	-	4.52	8.28	6.28
ES$_{total}$	0.84	0.56	1.43	0.79	0.55	1.18

DISCUSSION

To be able to quantify quality differences between L1 and L2 texts, the quality of both L1 and L2 texts must be expressed on the same scale. A method to achieve such a single scale and enable direct comparisons of L1 and L2 texts was, however, not yet available. In the present study, it was investigated whether a rating method with L1 and L2 benchmark essays and raters who are (near) native speakers of both L1 and L2 is suitable for direct comparisons of the rated quality of L1 and L2 essays. This was done by comparing the fit of five pre-specified nested models which all test the degree of measurement invariance across the L1 and L2 benchmark essays, and where each higher-level model imposes more restrictions of invariance across benchmark languages. It was found that there was no evidence against accepting the highly restrictive parallel_bm model, which means that ratings with L1 and L2 benchmarks may be assumed to be parallel tests. This indicates that the distribution of scores across L1 and L2 essays, as allocated by the raters, is similar, regardless of the language of the benchmark essay. In other words, there is no evidence to assume that the raters were unable to apply rating standards equally across languages. In addition, the raters performed their ratings with adequate levels of reliability. Therefore, the ratings of L1 and L2 essays are directly comparable and the quality difference (in terms of the allocated ratings) can be expressed. It was found that L1 essays received, on average, significantly higher ratings than L2 essays, indicating that the L1 essays are, in this specific population at least, of higher quality than the L2 essays.

The selection of raters who are (near) native users of L1 and L2 is essential to the creation of parallel ratings. Although the raters' Dutch (L1) and English (L2) language proficiency was not measured objectively in the present study and a minimum level of language proficiency can therefore not be specified, it seems possible to list some features which all raters held in common. In the present study, raters who were capable of performing parallel ratings with L1 and L2 benchmarks answered the following criteria: a) they used L1 and L2 on a daily basis (L2 mostly for professional communication), b) they had jobs which involved dealing with texts and were as such familiar with text conventions in both languages c) they had had an academic (linguistic) education, d) they had at least eighteen years of experience with both languages. Most of them were probably slightly more proficient in Dutch (the present study's L1) than in English (the present study's L2). Apparently, this minor imbalance was unproblematic in the creation of parallel rating scales.

While estimating the effect of language (i.e. L1 or L2) on the rated quality of essays, relatively large participant- and task-related score variances were found, in addition to large residual variance. Although the variance between different assignments (s^2 *(topic)*) is smaller than participant (s^2 *(participant)*) and error variances (s^2 *(residu)*), it is still quite large. This large between-assignment-variance, which is all the more striking as the tasks in the present study are highly similar, indicates that measurements with few assignments are unreliable representations of writing skill. Based on the estimates in table 5, a measurement of, for example, global quality with only one assignment (topic) has a reliability of only .33 if an L1 benchmark is used and .39 if an L2 benchmark is used. If global quality is assessed with four assignments per writer, the reliabilities of the obtained scores are .66 (L1 benchmark) and .72 (L2 benchmark). If eight assignments are used per writer, the reliabilities with which differences between writers can be established are .80 (L1 benchmark) and .83 (L2 benchmark). In other words, measurements of writing skill with few tasks are likely to be non-representative. Van den Bergh (1988b) argues that writing assessments on the basis of single tasks might basically be regarded as single-item-tests, which do not allow for generalizations about an individual's writing ability. While the use of as many participants as possible is common practice in writing research, and rightly so, measurements of writing skill are still quite often conducted using just one writing assignment per condition. Clearly, this is not advisable (cf. Gebril, 2009; Schoonen, 2005), as it probably supplies insufficient information concerning an individual's writing ability, and because it does not allow for establishing whether any effects found are really effects of the specified condition, or just effects due to task (Rijlaarsdam et al., 2011). For example, differences between L1 and L2 writing can never be established if only one writing task is used per language (cf. Van Weijen, 2009). After all, any differences found between the two languages might actually be due to task. Hence, the large language effect found in the present study could only be uncovered because multiple tasks were used in both L1 and L2. The minimum number of tasks to be used for a reliable assessment depends, among other things, on the applied rating method, number of raters, test population, benchmark language and task type (cf. Coffman, 1966; Schoonen, 2005; Van den Bergh, 1988b). Residual variance is, in all cases, larger than the assignment-related and participant-related variance, indicating that a large part of the score differences cannot be attributed to differences between assignments or participants. Residual variance includes the interaction between participant and assignment (e.g. some assignments are more difficult than others, but this added difficulty will be different for different participants).

The jury reliabilities were substantially higher than individual rater reliabilities. In other words, the reliability of the measurement increases quite drastically if the ratings are performed by multiple raters. This finding reinforces a point which has been made by many (Gebril, 2009; Raymond, 1982; Schoonen, 2005; Van den Bergh & Eiting, 1989; Weigle, 2002), namely that assessments by single raters are insufficient. Just as single-item tests cannot provide reliable measurements, so single-rater assessments cannot either.

The correlations between the three different rating criteria (global quality, structure and language) are strong. This is not uncommon in essay assessment studies (cf. De Glopper, 1988; Van den Bergh, 1988a). In addition, the reported correlations are disattenuated correlations. These are always stronger than correlations uncorrected for measurement error, which are usually reported. Nevertheless, the ratings of the three criteria do not seem to be fully distinguishable. There are probably two explanations for this. One is that, to some extent, criteria may truly coincide (cf. Bae & Bachman, 2010). As such, global quality includes the two other rating criteria, structure and language. The strong correlations between global quality and structure and global quality and language should therefore probably not be seen as an indication of major validity problems with the applied rating criteria. After all, if criteria (partly) coincide, they are expected to correlate. On the other hand, the perfect correlation between global quality and structure in the present study probably exceeds the expected correlation and indeed raises questions about the validity of these two rating criteria. Were the raters able to look beyond the structure of an essay (which of course impacts its content and persuasive power) while rating its global quality? The second explanation for the high correlations is therefore the possible occurrence of halo effects. Raters' global impressions of an essay may have spilled over to the ratings of single aspects (structure and language), or vice versa, even though care was taken in the present study to minimize this possibility (e.g. ratings of different criteria were carried out separately). This may be the reason why structure and language – the two rating criteria which are most expected to be independent categories – correlate relatively substantially, although to a lesser degree than they correlate with global quality. In any case, the strength of the correlations is not directly relevant for the question whether ratings with L1 and L2 benchmarks are parallel tests. The correlations are only required to be similar across benchmark languages.

In the present study, the L2 (English) was a language from a culture which is relatively similar to the L1 (Dutch) culture, which means that the ideas about what constitutes a good piece of writing are unlikely to differ greatly. If the L1 and L2 cultures are less similar to each other, it might be harder for the raters to rate L1 essays relative to L2 benchmarks and vice versa, even if they are (near) native speakers of both languages. In those cases, the benchmark essay may be so different from the essay to be rated, due to cultural standards, that comparisons become extremely hard, which will have a negative impact on rater reliability. In addition, the allocation of scores across L1 and L2 essays is, in such cases, likely to be different for the two benchmark languages. For example, the L2 essay scores may have a far smaller range than the L1 essay scores, if an L1 benchmark is used, whereas the range of L1 essays is likely to be smaller than the range of L2 essays if an L2 benchmark is used. In other words, true and error score variances might not be invariant across benchmark languages if the L1 and L2 cultures are very dissimilar, so that ratings with L1 and L2 benchmarks might in such cases not be parallel tests. Whether parallelism can be achieved in situations with different combinations of first and second languages should be tested for every new occasion in which it is applied.

Thus far, it had not been possible to express the quality of L1 and L2 essays on a single scale. In the present study ratings of L1 and L2 essays conducted by raters who are (near) native or bilingual speakers of both L1 and L2 with benchmark essays in both L1 and L2 were found to be parallel tests. Therefore, comparison of L1 and L2 essay scores is allowed and quality differences between the two languages can be expressed. While a test of parallelism will need to be conducted in every new instance where this procedure is used, the rating procedure applied in the present study seems a promising method for further comparisons of L1 and L2 writing, for example in different populations. In the population investigated in the present study, L2 writing quality was homogeneously lower than L1 writing quality.

CHAPTER 3

QUALITY DIFFERENCES BETWEEN L1 AND L2 WRITING. COMPARING L1 AND L2 TEXTS AND WRITING PROCESSES

Abstract [4]

When unaccomplished L2 learners write L1 and L2 texts, L2 text quality is often lower than L1 text quality, in terms of language use as well as in terms of content and organization (cf. chapter 2). This study sets out to investigate whether the underlying writing processes explain deterioration of writing performance in L2. Twenty fourteen- and fifteen-year-olds wrote four essays in L1 (Dutch) and four essays in L2 (English, taught as a foreign language). The quality of each text was rated by three expert raters (independently of each other). Students' writing processes were registered by means of think aloud procedures, combined with keystroke logging. This study unites three methodological advantages, which have hitherto not been united in one study. First, multiple tasks are used per language, so that we can separate language effects from task effects. Second, writing processes were analyzed temporally: the moment at which cognitive activities (such as planning, formulating, and revising) occur during writing (i.e. at the start or toward the end of task execution) reflects the quality of the writing process. Third, L1 and L2 text quality are expressed on the same scale. This set-up allows us to make a direct comparison of the L1 and L2 relations between the writing process and text quality. Results show that for the age group under study, effective L2 writing processes are different from effective L1 writing processes. L2 text quality increases if students distribute their attention to cognitive activities differently across task execution during L2 writing than during L1 writing.

[4] An adapted version of this chapter has been submitted for publication as: Tillema, M., Van den Bergh, H, Rijlaarsdam, G., & Sanders, T. (submitted). Quality differences between L1 and L2 texts and writing processes, and the mediating role of linguistic proficiency. An empirical study.

When students write in a second language (L2), these texts are often of substantially lower quality than texts they write in their first language (L1), in terms of language use as well as in terms of content and organization (Silva, 1993; see also chapter 2). This quality difference is generally assumed to reflect the added difficulty of L2 writing (Silva, 1993; Roca de Larios et al., 2008; Schoonen et al., 2003; Thorson, 2000; Van Weijen et al., 2008), particularly for unaccomplished language learners who have not yet fully mastered the L2.

To explain this deteriorated quality in L2, a number of researchers have investigated and compared L1 and L2 writing processes. Quite often, these L1/L2 process comparisons are based on one task per writer per language (Chenoweth & Hayes, 2001; Roca de Larios, Manchón & Murphy, 1996; Silva, 1993[5]; Uzawa, 1996). As a result, interpretations of these studies' results are problematic, as many researchers (cf. Rijlaarsdam et al., 2011; Van den Bergh et al., 2009; Van Weijen, 2009) have demonstrated that there can be enormous differences between individual writing processes in one language due to task, even if tasks only differ in terms of topic (cf. Rijlaarsdam et al., 2011; Van den Bergh et al., 2009; Van Weijen, 2009). If only one task is used per language, it is therefore impossible to tell if any differences between L1 and L2 processes are indeed language effects, or merely task effects. Van Weijen (2009, p. 94) demonstrated that, on average, over 50% of the variance within (18-year-old) writers is due to task in L1, and over 35% in L2. This implies that L1/L2 comparisons of writing processes based on single tasks per language overestimate the difference due to language.

Stevenson, Schoonen and De Glopper (2006) investigated 13- and 14-year-old students' revision processes in L1 and L2, using two tasks per language per writer. They analyzed revision processes in terms of their average frequency of occurrence during the writing process and found few differences between L1 and L2 writing: writers made more linguistic revisions (i.e. revisions below clause level and revisions of language or typing) in L2 than in L1, but for 'higher order' revisions (i.e. revisions above clause level and revisions of content) no difference between L1 and L2 was found. Note, however, that this result does not rule out the possibility that the L1 and the L2 condition require different *distributions* of revision activities across task execution. Even if the numbers of revision activities do not differ between L1 and L2 writing, it might still be possible that writers revise at different moments during writing in L1 and L2.

Over the last two decades, researchers have increasingly acknowledged and demonstrated that modelling temporal distributions of cognitive activities across

[5] Silva (1993) is an overview study. The included studies on L1 and L2 writing processes only involve multiple tasks per language in a small minority of cases.

the writing process is a more sensitive and valid reflection of the quality of processing than an analysis in terms of average frequencies (Breetvelt et al., 1994; Leijten & Van Waes, 2006; Olive et al., 2008; Roca de Larios et al., 2011; Rijlaarsdam & Van den Bergh, 1996; Rijlaarsdam et al., 2011; Van den Bergh & Rijlaarsdam, 1996; Van den Bergh et al., 2009; Van der Hoeven, 1997). In other words: quality of processing during writing is better reflected by analyzing *the moment at which* cognitive activities (such as planning, formulating and revising) are applied than by analyzing *how often* these activities occur during writing. Rijlaarsdam and Van den Bergh (1996) and Van den Bergh and Rijlaarsdam (1996), for example, demonstrated that the occurrence of cognitive activities varies across task execution. Structuring activities, for example, are on average more likely to occur a short while after the start and also towards the end of task execution, but less likely to occur during middle stages of the writing process. In addition, these researchers found that the degree to which cognitive activities contribute to text quality also varies across the writing process, e.g. structuring activities were more effective when they occurred during early stages of task execution (i.e. the correlation between structuring and text quality is at its highest at the start of the writing process) and less effective when they occurred towards the end of task execution (Rijlaarsdam & Van den Bergh, 1996). The researchers therefore advocated a temporal analysis of activities over the writing process: analyses of cognitive processing during writing should take the moment(s) at which cognitive activities occur into account.

The temporal approach was implemented by Van Weijen (2009) to study 18-year-olds' process variation during L1 and L2 writing, using four tasks per language. She found that the average temporal distributions of cognitive activities (i.e. reading, planning, generating, and formulating) in L2 are quite similar to the L1 distributions. Both in the L1 and the L2, for example, formulating activities are unlikely to occur at the start and end of the writing process, but quite likely to occur just before the writing process is halfway finished. So: the similarity between L1 and L2 writing processes found by Stevenson et al. (2006) in terms of frequencies of cognitive activities (revisions) was confirmed by Van Weijen (2009), who investigated average temporal distributions of cognitive activities.

In addition to an L1-L2 comparison of processes, Van Weijen (2009) also investigated whether the contribution of writing processes to text quality is similar in L2 and L1 situations. The results indicated a different effect of writing processes on L1 and L2 text quality: the most effective distributions of cognitive activities across the writing process were different in L1 and L2. For example, whereas the correlation between formulating activities and text quality is lowest at the start and end of the writing process and highest after about a quarter of the writing process has been finished in L1, this correlation is lowest at the start and highest at the end of the writing process in L2 (Van Weijen, 2009, p. 95). Van Weijen (2009) suggests that the added language difficulty during L2 writing tasks makes L2 writing cognitively more demanding. Writers may need to adapt to these extra demands in L2 writing by orchestrating their writing processes differently during L2 writing tasks than during L1 writing tasks.

In Van Weijen's (2009) study, processes and relations between process execution and text quality were modeled for L1 and L2 writing separately, because L1 and L2 text quality scores were not expressed on the same scale and therefore not directly comparable. In fact, this is usually not the case: comparing text in two (or more) different languages is problematic, as text quality is generally not assessed in a similar manner across languages. For example, raters may apply different quality standards to different languages or quality standards may not be applicable to both languages). Therefore, L1 and L2 text scores are in general assumed not to be directly comparable, as they are usually not expressed on parallel scales. Should a direct comparison of L1 and L2 text scores be desired, a rating procedure is needed which is expected to allow comparability. Subsequent to the implementation of such a procedure, a statistical check is needed to establish whether parallel ratings of L1 and L2 text quality were indeed achieved. Van Weijen (2009, p. 86-87) set up such a procedure, involving translations of L1 essays into L2. However, the procedure did not result in a single scale on which to place both the L1 and L2 text scores and compare them directly. Therefore, a separate analysis of L1 and L2 writing was necessary in Van Weijen's (2009) work.

The drawback of analyzing relations between process execution and text quality for L1 and L2 writing independently is that it is not possible to directly compare the contributions of cognitive activities to text quality across L1 and L2 writing. A question such as "do planning activities contribute more strongly to text quality during initial stages of task execution during L2 writing than during L1 writing?" cannot be answered, for example. However, obtaining an answer to such a question is very important, as it means that we can establish whether and, if so, how L2 writing processes should differ from L1 writing processes, in order to write L2 texts of as high a quality as possible. The set-up of the present study allows for a direct comparison of L1 and L2 relations between cognitive activities and text quality across the writing process. It brings together a number of methodological advantages which have hitherto not been united in one study. First, the problem of incomparability of L1 and L2 text scores has been overcome: the ratings of L1 and L2 essays are expressed on a single scale (as was established in chapter 2). This allows us to compare the contributions of cognitive activities to text quality across L1 and L2 writing, as they can be analyzed in one and the same model. Second, multiple tasks are used per writer and per language, so that we can separate language effects from task effects. Third, the writing processes, and relations between writing processes and text quality, are analyzed in terms of the temporal distributions of cognitive activities, as such temporal distributions are valid reflections of the quality of the writing process.

The present study, then, is designed to investigate whether the contribution of cognitive activities to text quality at various stages of the writing process is different during L1 and L2 writing. To answer this question, it is necessary to first find out if L2 writing processes are different from L1 writing processes, as this influences the interpretation of a cognitive activity's contribution to text quality in L1 and L2. So while research question 2 is the main question of the present study, the research questions in chronological order are:

1. Do the temporal distributions of cognitive activities across the L2 writing processes differ from their temporal distributions across the L1 writing process?

2. Is the contribution of cognitive activities (at various stages of the writing process) to text quality different during L1 and L2 writing?

METHOD

Participants

The participants were fourteen- and fifteen-year-old students (N = 20; 10 female and 10 male). They were recruited from three different third-year-forms at the same school for pre-university secondary education by means of a call for volunteers, which was distributed by their Dutch language teacher. All participants were native speakers of Dutch. They received a small financial reward for their participation. Parental consent was obtained.

Instruments and procedures

Writing tasks

The students completed eight writing tasks. They wrote four argumentative essays in L1 (Dutch), and four argumentative essays in L2 (English), on topics such as 'camera surveillance in inner city areas' or 'legalisation of soft drugs'. Multiple tasks were used per language to disentangle task effects and language effects. After all, if only one task were used per language, it would be impossible to know if any differences which are found are due to task or due to language (Van den Bergh et al., 2009; Van Weijen, 2009).

All eight writing assignment were similar in terms of audience (peers), medium (a school-related magazine for secondary school students) and purpose (to convince the readers of your point of view), and differed only in terms of topic. The essays had to be about half a page (A4 format) in length (which is about 250 to 300 words). The assignments were tested with third year students of pre-university secondary education during a pilot study in 2005. They were also successfully used in previous studies (Van Weijen et al., 2008; Van Weijen, 2009). An example of an assignment can be found in Appendix A.

The available time for each essay was approximately thirty minutes, although participants were allowed to go on longer if they felt that their essays were not finished yet. No participant used more than forty minutes. The students completed the essays during two separate days. Between the four tasks completed per day, participants were given a break of about ten to fifteen minutes. No difference in text quality was established due to the order in which the four essays were completed (per day). For all comparisons (e.g. 1st essay of the day vs. 2nd essay of the day; 2nd essay of the day vs. 3rd essay of the day, et cetera): $\chi^2 \leq 1.36$, $df = 1$, $p \geq .24$.

To avoid sequence effects and to control for effects of topic, several measures were taken. It has been established that topic can greatly influence the quality of writing (Godshalk et al., 1966; Rijlaarsdam et al., 2011; Schoonen, 2005; Van den Bergh et al., 2009; Van Weijen, 2009). Therefore, to disentangle language effects from topic effects, the topics were systematically balanced across languages, so that each topic occurred in both L1 and L2. So, writer 1 wrote on topics 1, 2, 3 and 4 in L1 and on topics 5, 6, 7 and 8 in L2, whereas writer 2 wrote on topics 2, 3, 4 and 5 in L1 and on topics 1, 6, 7, and 8 in L2, and so forth. The order in which L1 and L2 essays were written was also balanced across participants: ten students first completed four L1 essays, and then wrote four L2 essays; the other ten students first completed the L2 essays, and then wrote the L1 essays.

The students wrote the essays on a computer using Microsoft Word. They were all very familiar with using MS Word. The students had to think aloud during the process of task execution. If they fell silent, the test leader prompted them to continue thinking aloud by a neutral remark: “aloud, please” (in Dutch). The participants practiced thinking aloud before writing their essays by means of a two-line writing assignment and a short mathematical puzzle. All writing sessions were audio- and video-taped and recorded by means of keystroke logging (Inputlog: Leijten & Van Waes, 2006).

Coding process data

All think aloud data were transcribed and segmented, and completed by the Inputlog recordings (system adapted from Van Weijen, 2009). A new segment in the protocols reflected a switch to a different cognitive activity within a participant's writing session (Van den Bergh & Rijlaarsdam, 2001). All segments were coded according to a coding scheme (adapted from Breetvelt et al., 1994). The coding scheme (see table 1) consists of fourteen categories. The 'Revising' category involved subcodes. Revisions were subcoded as 'automated corrections' when they involved corrections of typographic errors. They were typically errors which are made as a result of the use of a keyboard, which seem to be corrected almost automatically. An example would be: a writer types 'almots', and immediately corrects this error

Table 1. Coding categories in the coding scheme

	Coding category	**Description**	**Example**
READING THE ASSIGNMENT	Reading the instruction text and documentation	Reading (part of) the task instructions	"Write an essay in which you..."
PROCESS PLANNING	Monitoring	Verbalizations indicating a steering capacity which governs the writing process, mostly self-instructions	"I'm going to read what I've written so far."
	Metacomments	Evaluations of a student's own writing process	"I should have made an outline of the text before I started."
CONTENT PLANNING	Goal setting	The formulation of goals which the text has to satisfy	"The text should be convincing."
	Generating	Generating ideas for content or form	"Something about the disadvantages of camera surveillance..."
	Structuring	Evaluating and arranging ideas	"Something about adults? No, that's not relevant."

Table 1 (continued). Coding categories in the coding scheme

FORMULATING	Text production	Production of new text	"Camera surveillance invades people's privacy." (Verbalizations usually occur parallel to the activity of typing.)
READING OWN TEXT	Reading produced text	Reading (part of) the produced text, at any given moment during the writing process	"Surveillance cameras do not increase public security."
EVALUATING OWN TEXT	Evaluating produced text	Evaluation of produced text	"The largest part is about drawbacks."
REVISING	Conceptual revisions	Making changes (at word, sentence or text level) to the text produced so far	Moving a set of sentences from the body of the text to the introduction.
	Automated corrections	Corrections of typographic errors due to keyboard use. *(Not included in the analyses.)*	Writer types 'almots', deletes 'ts' by means of 'backspace', en then types 'st'. (Mostly no (explicit) verbalization of the correction.)
OTHER	Pauses	Silence or interjection	"eeeerrr"
	Interaction with test leader	Interaction between test taker and test leader	"Could I open a window?"
	Physical activity	Physical activity	Taking a sip of tea.
	Navigation	Moving through the document: arrow buttons or mouse movements	Moving the cursor some lines back.

by giving two backspaces and typing 'st', so that it now reads 'almost'. All revisions which were not 'automated corrections' were subcoded as 'conceptual revisions': they involve alterations which are made by the writer in a non-automated way. That is, they involve actual alterations at the level of spelling or content. In the remainder of this article, we will use the term 'revision' or 'revising' only if we are referring to conceptual revisions. One randomly chosen think aloud protocol (number of segments = 518) was coded by two researchers. The intercoder agreement (Kappa) was satisfactory (.85).

As we are interested in complete writing processes, all cognitive activities in table 1 are analyzed, except for the activities listed under 'Other'. The reason for excluding the 'Other' activities is that they are either too diffuse (Pausing and Navigating can reflect all kinds of cognitive processing, or none at all) or conceptually irrelevant and highly infrequent (Interaction With Test Leader and Physical Activity).

Table 2 shows part of a protocol and illustrates how the think aloud data en the Inputlog data were integrated. The protocols consisted of seven columns. The column labeled 'Reading' was used for indicating if any reading activities (Reading the Assignment - RA – or Reading Own Text – ROT) were taking place. The 'Verbalizations' column contained everything which was said out loud by the student, except interjections, such as "uhm". All information in the column labelled 'Typing' are derived from Inputlog. There were three categories in this column, namely the production of new text, revisions (indicated by Inputlog as [BS] for backspace or [DEL] for if the delete button was pressed), and navigation (by means of mouse movements or arrow buttons). The 'Pausing' column contained all

Table 2. Part of a completed protocol

Segment	Reading	Verbalizations	Typing	Pausing	Other	Code
17		even een titel erboven (*I'll insert a title)*				Monitor
18		having	Having childere			Text production
19		chil	[BS 1] [BS 1] [BS 1]			revision (automated correction)
20		children yes or no	ren, yes or no?			text production
21	ROT	having children yes or no				Reading own text

silences and interjections. The column labeled 'Other' mostly contained descriptions of physical activities, for example 'takes a sip of his drink'. One row is one protocol segment. As such, this transcription method allows for parallel actions. Text production and verbalizations, for example, often occur simultaneously. The 'Typing' column would contain the text production as registered by Inputlog. The codes in the last column were in reality numbers. Code 01, would stand for 'reading the assignment', for example, and code 02 would stand for monitoring. English translations of Dutch verbalizations are given in italics.

Due to technical deficits (defective videotapes or unrecorded keystroke loggings), sometimes not all eight writing tasks carried out per participant were available for the analyses of the writing processes (although all essays were available for the text quality ratings described in the next paragraph). Table 3 shows the exact number of tasks included in the analysis per language condition and per participant. It is clear that, although the measurement sometimes involves less than four tasks per language condition per student, the L1 and L2 writing processes of all students have been measured using multiple tasks, which enhances the measurement's reliability.

Table 3. Number of tasks included in the analysis

Participants	L1	L2
2, 3, 4, 5, 7, 8, 10, 11, 13, 15,	4	4
9, 12, 14, 16, 17, 18, 20	4	3
6, 19	3	3
1	2	4

Assessing text quality

Eight experienced raters were involved in this study, all of whom rated overlapping samples of 120 out of the 160 available essays. All raters performed their ratings independently of each other. The raters who rate the same set of essays form a jury. The ratings were conducted with an average jury reliability of .75. The raters judged the essays on three criteria of text quality: global quality, structure and language. To minimize the possibility that the ratings for one criterion affect the ratings on the other criteria, the ratings were carried out in three different rounds (on different days, one round per rating criterion). During each new round, the ratings for previous criteria were no longer available to the raters. This procedure notwithstanding, the ratings on global quality, structure and language use were found to be strongly correlated (global quality – structure, $r = 1$; global quality –

language use, *r* = .89; structure – language use, *r* = .75)[6], which is not uncommon in essay assessment studies (cf. De Glopper, 1988; Van den Bergh, 1988a). Therefore, the analyses were carried out using text scores (per writer, per task) which were the averages of the scores assigned to the three criteria (by three raters).

The raters applied rating standards equally strict (or lenient) to essays in both languages. Direct comparisons between L1 and L2 essays scores are therefore allowed. (For a detailed explanation of the procedure by which this result was achieved: see Appendix D).

Analyses

Analyses of writing processes should take 'time during the writing process' into account as an explanatory variable to operationalize the writing process in a valid and sensitive way (Rijlaarsdam & Van den Bergh, 1996; Van den Bergh & Rijlaarsdam, 1996). In the present study, this was accomplished by splitting each protocol into five equally long episodes in terms of numbers of segments (cf. Breetvelt et al., 1994; Roca de Larios et al., 2008; Van den Bergh & Rijlaarsdam, 1996; see also chapter 5). A protocol of 330 segments, for example, would be analyzed as five episodes consisting of 66 segments each. For each cognitive activity under analysis (i.e. reading the assignment, process planning, content planning, text production, reading own text, evaluating own text, and revising) its percentage of occurrence relative to the total number of segments was calculated for each episode. For instance, if an episode consisted of 80 segments (which means that the entire writing process consisted of 400 segments), and 10 of these segments were coded as 'reading the assignment', then the percentage for reading the assignment in that episode would be 12.5. This way, the percentage with which a cognitive activity occurs is allowed to vary across the writing process.

To answer research question 1, a regression model has been applied to model the occurrence of each cognitive activity as a function of *episode* (the percentage of occurrence may be different at different moments during the writing process) and of *language* (the percentage of occurrence may be different during L1 and L2 writing). A multilevel regression model was used, as our data are hierarchically organized, i.e. episodes are nested within tasks and participants (Van den Bergh et al., 2009). As the effect of language on the occurrence of a cognitive activity may be different during the various episodes in the writing process, an interaction variable *episode*language* was also entered as a predictor variable.

[6] These are disattenuated correlations, i.e. correlations between true scores, ridded of measurement unreliability.

To answer research question 2, the effect of variations in the occurrence of a cognitive activity across task execution on text quality was modeled by entering each episode as a predictor variable in multilevel regression analysis. This, again, was done for each cognitive activity separately. Text quality was described as a function of each of the five episodes, i.e. TQ=*f*(epi1, epi 2, epi 3, epi 4, epi 5). This function *f* need not be identical for all individuals *i*, or tasks *j*: $TQ_{(ij)}=f_{ij}(episode1, episode\ 2, episode\ 3, episode\ 4, episode\ 5)$.

So far, the model describes the average effect of the percentage of occurrence of a specific cognitive activity, per episode, on text quality, regardless of the language in which the writing took place. However, the model was extended to model the deviations if it concerns an L2 writing process. That is: five new variables were created and entered into the equation as predictors. These variables were the percentages of the cognitive activity in question per episode, differing from the already entered episode variables in that these new variables only exist if the text was written in L2:

$$\begin{aligned} TQ_{(ij)} &= \beta 0_{ij} + \beta 1 * epi1_{ij} + \beta 2 * epi2_{ij} + \beta 2 * epi2_{ij} + \beta 3 * epi3_{ij} + \beta 4 * epi4_{ij} + \beta 5 * epi5_{ij} \\ &+ L2_{ij} * (\beta 6 * epi1_{ij} + \beta 7 * epi2_{ij} + \beta 8 * epi3_{ij} + \beta 9 * epi4_{ij} + \beta 10 * epi5_{ij}) + e_{(ij)} + u_i + v_j \end{aligned}$$

In effect, the first half of the model (*β1- β5*) now describes the effect of how often the cognitive activity occurs per episode on text quality in L1, while the second half of the model (*β6- β10*) describes whether there is a significant effect change (relative to the L1 effect) if writing occurs in L2. However, if a significant effect change occurs, this does not automatically mean that the cognitive activity significantly contributes to text quality in L2. After all, if we find that applying a cognitive activity during a specific episode has a (positive or negative) effect on text quality in L1, and if we find that this positive relation significantly changes (e.g. becomes smaller) if writing occurs in L2, it does not automatically mean that the activity is still related to text quality in L2. Therefore, the significance level of the L2 estimates (e.g. the L2 estimate for a cognitive activity in episode 1 equals *β1+ β6*) was tested. The constructed model, then, is a solution for testing different kinds of relationships at once. As they are analyzed in a single model, the L1 and L2 regression weights – indicating the contribution of a cognitive activity to text quality at a given episode – can be directly compared.

RESULTS

Writing process differences in L1 and L2 (research question 1)

Table 4 gives an overview of the effects of episode and language on the occurrence of each of the analyzed cognitive activities. Parameter estimates are given for each

Table 4. Effects of episode and language on the occurrence (in percentages) of seven cognitive activities; r = correlation between predicted and observed values

	Main effects		**Interaction effect**		
Cognitive activity	**Episode**	**Language**	**Episode*language**		***r***
Reading Assignment	Epi1: 10.74 Epi2: 5.33 Epi3: 5.67 Epi4: 4.14 Epi5: 2.06				.64
Process Planning			L1 Epi1: 5.65 Epi2: 2.92 Epi3: 2.77 Epi4: 1.44 Epi5: 1.96	L2 Epi1: 4.60 Epi2: 2.74 Epi3: 1.95 Epi4: 1.29 Epi5: 3.12	.60
Content Planning	Epi1: 4.68 Epi2: 3.52 Epi3: 3.36 Epi4: 2.84 Epi5: 2.23	L1: -1.02			.65
Formulating			L1 Epi1: 34.28 Epi2: 39.24 Epi3: 39.70 Epi4: 40.33 Epi5: 35.50	L2 Epi1: 29.31 Epi2: 36.76 Epi3: 38.26 Epi4: 38.41 Epi5: 36.00	.67
Reading Own Text					.73
Evaluating Own Text	Epi1: 0.30 Epi2: 0.61 Epi3: 0.53 Epi4: 0.75 Epi5: 1.45				.47
Revising					.52

established effect. This table[7], then, reports differences between L1 and L2 writing processes. (For an overview of effects per predictor variable, see Appendix E). To assess the degree to which these parameters explain the total variance in the data, we correlated the values as predicted by the models and the observed values. These correlations are also presented in table 4. Their values are (moderately) satisfactory. It seems that episode and language explain a substantial part of the variance, but that some variance remains unexplained.

No effect of language was found for Reading the Assignment, Reading Own Text, Evaluating Own Text, and Revising (i.e. conceptual revisions). For Reading Own Text, no effect of episode exists either, meaning that we have to assume that this activity in general occurs equally often in all episodes and in both languages (i.e. we cannot assume that there is a difference between languages and episodes). This also applies to Revising. For Reading the Assignment and for Evaluating Own Text, a main effect of episode was found. The parameter estimates indicate that the occurrence (in percentages) of Reading the Assignment generally decreases as the writing process progresses: it is most likely to occur (i.e. in 10.74% of the segments) during episode 1 and least likely to occur (i.e. in 2.06 % of the segments) during episode 5. As no main effect of language existed for this activity, the described distribution and amounts hold for Reading the Assignment in both L1 and L2. The occurrence of Evaluating Own Text generally increases as the writing process progresses: from 0.3% at the start of task execution (i.e. episode 1) to 1.45% at the end of task execution. This, again, holds for both L1 and L2.

A main effect of language was found for Content Planning. In addition, a main effect of episode was established for Content Planning. The probability that Content Planning occurs generally decreases (from 4.68% to 2.23%) as the writing process progresses, but during all five episodes, the amount of occurrences of Content Planning is (1.02%) lower in L1 than in L2.

An interaction effect of language and episode was established for Process Planning and for Formulating, meaning that the effect of language is different for different episodes. The parameter estimates indicate that, during episodes 1 through 4, Process Planning occurs more during L1 tasks, on average, than during L2 tasks. In episode 5, Process Planning occurs more during L2 tasks than during L1 tasks. Formulating, too, occurs more during L1 tasks than during L2 tasks in episodes 1 through 4, but less during L1 tasks than during L2 tasks in episode 5.

[7] 'Pausing' and 'automated corrections' were relatively frequently occurring activities. This explains why the accumulated percentages (per episode) for the cognitive activities presented in table 4 do not approximate 100.

Relations between text quality and writing processes in L1 and L2 (research question 2)

Table 5 features descriptive information about the text quality scores. L2 texts scored, on average, 134.1 – 93.5 = 40.6 points lower than L1 texts. Given the standard deviation for both L1 and L2 texts, the overlap in text scores (i.e. high scoring L2 texts and low scoring L1 texts) was relatively small. Effects sizes (*Cohen's d* >1) indicate that, while score variations within languages are quite large in both L1 and L2 (though somewhat smaller in L2), there is a relatively large difference between L2 and L1 text scores (see chapter 2).

Table 5. Descriptive statistics for the L1 and L2 text scores; Means (M), standard deviations (SD), minimum and maximum scores/percentages (Min.; Max.)

	L1	**L2**
M	134.1	93.5
SD	34.2	27.8
Min.	73.6	30.3
Max.	257.8	185.0

We now know that there is a large text quality difference between L1 and L2 texts. We also know that Process Planning, Content Planning and Formulating are carried out differently during L2 writing than during L1 writing, but that no L2/L1 difference was found for Reading the Assignment, Reading Own Text, Evaluating Own Text and Revising. The next question to answer is whether relations between (distributions of) cognitive activities and text quality are different for L1 and L2 writing.

For each cognitive activity, the fit of three regression models (in which text quality is explained by the writing process) was assessed by comparing the respective -2 log likelihoods: a model in which only the effect of the mean occurrence during the entire writing process on text quality was included, i.e. an intercept-only model, a model with the five episodes as predictor variables, and a model with episodes, language (L1 vs. L2) as predictor variables. For all seven cognitive activities, the models with episodes and language as predictors were better fits to the data than the intercept-only models (χ^2 > 94.4, *df* = 10, *p* < .001) or the episodes-only models (χ^2 > 77.95, *df* = 5, *p* < .001). This means that, for all seven cognitive activities, the relation between writing process and text quality varies across the writing process (i.e. between episodes) as well as between languages. The next step now is to investigate the nature of these relations.

Table 6 gives the effects of the seven cognitive activities on text quality during different episodes of the writing process. It shows, per episode and for both L1 and L2 writing, whether any relations between variations in text quality and variations in the occurrence of a cognitive activity were found and, if so, if these effects are positive or negative (see Appendix F for parameter estimates). Table 6 tells us that the relations between text quality and cognitive activities are dissimilar for L1 and L2 writing for all of the seven cognitive activities. They are dissimilar in two ways.

First, cognitive activities are generally related to text quality (whether positively or negatively) at different episodes during the writing process in L1 and L2. For example, a relation between the occurrence of Reading the Assignment and text quality was established in episode 2 during L1 writing. In L2, however, the

Table 6. Effects of cognitive activities on text quality per episode
L1 = effect of activity for L1 writing (+ = positive effect; - = negative effect)
L2 = effect of activity for L2 writing (+ = positive effect; - = negative effect)
Empty boxes indicate that no significant effects could be established (p > .05)

		Episode 1	**Episode 2**	**Episode 3**	**Episode 4**	**Episode 5**
Reading Assignment	L1		+			
	L2	-	-			
Process Planning	L1			+	-	+
	L2	-			-	-
Content Planning	L1	+		+		
	L2				-	
Formulating	L1					
	L2					-
Reading Own Text	L1				+	+
	L2					
Evaluating Own Text	L1		+		+	+
	L2		-	+	-	-
Revising	L1			+		+
	L2					

percentage of occurrence of Reading the Assignment is related to text quality in episode 2, but also in episode 1. Content Planning is related to text quality in episodes 1 and 3 during L1 writing, but in episode 4 during L2 writing. In short, the episodes in which variation in the occurrence of the activity is related to variation in text quality are always (sometimes partly, sometimes completely) different during L1 and L2 writing for all seven cognitive activities.

Second, where variations in the occurrence of cognitive activities are related to variations in text quality in the same episodes during L1 and L2 writing, the direction of the relation between cognitive activities and text quality regularly seems to be dissimilar for L1 and L2 writing. For instance: the relation between the occurrence of Reading the Assignment and text quality in episode 2 is positive in L1 but negative in L2. Or: the more Reading the Assignment occurs in episode 2 during L1 writing – relative to the average proportion with which Reading the Assignment occurs in L1 – the higher the L1 text quality score. And: the more Reading the Assignment occurs in episode 2 during L2 writing – relative to the average proportion with which Reading the Assignment occurs in L2 – the lower the L2 text quality score. The same seems to apply to Process Planning in episode 5, and to Evaluating Own Text in episodes 2, 4 and 5.

However, the interpretation of these directions depends on whether the cognitive activity in question was carried out differently during L1 and L2 writing (see table 4). In fact, for those cognitive activities which are applied differently during L1 and L2 writing, we cannot compare the directions of the established relations (table 6) in L1 and L2 without taking into account the writing process information (table 4). The reason for this is that 'applying an activity more (or less) than average' during a certain episode (which often results in a higher or lower text quality, as indicated by the pluses and minuses in table 6) means something different in L1 and L2 if the average occurrences in L1 and L2 are (for that specific episode) not the same. This applies for Process Planning, Content Planning, and Formulating. After all, for these three activities an effect of language was established: their average occurrence was different in L1 and L2 (in every episode). However, the occurrence of Content Planning and Formulating is never related to text quality in the same episodes during L1 and L2 writing. A comparison of L1/L2 differences in directions is therefore only necessary for Process Planning. This comparison is described in the section below.

Comparing L1 and L2 directions of relations between text quality and Process Planning
Variations in the occurrence of Process Planning are related to variations in text quality during episode 4 (negative in L1 and L2) and episode 5 (positive in L1, negative in L2) in both languages. However, the L1 and L2 relations between text

quality and Process Planning are not directly comparable, as the average occurrence of Process Planning was different during L1 and L2 writing, in every episode (see table 4). The directions (positive or negative) of the relations between text quality and Process Planning in L1 and L2 are interpreted differently according to the percentage with which Process Planning occurs in each of the episodes during L1 and L2 writing.

In episode 4, the relation between the occurrence of Process Planning and text quality is negative in both L1 and L2, according to table 6. However, Process Planning occurs less in L2 (1.29%) than in L1 (1.44%). So if, for some hypothetical student, Process Planning activities were to occur in 1.35% of the segments in both L1 and L2, this would mean 'doing more of it than average' in L2, which is related to a decrease of L2 text quality, while it would mean 'doing less of it than average' in L1, which is related to an increase of text quality. So, whereas the directions of the relations between Process Planning and text quality in episode 4 seem similar in L1 and L2 (negative), applying Process Planning activities in equal amounts during L1 and L2 writing has a different effect on text quality in L1 and L2 if the amount lies between 1.29% and 1.44%. For percentages over 1.44% or under 1.29%, the relation between text quality and Process Planning is similar for both L1 and L2. After all, any percentage over 1.44% is more than average in both languages, which has a negative effect on text quality in L1 and L2 (and vice versa for any percentage below 1.29%). So, while table 6 seems to indicate that there is a negative relation in L1 and L2, a positive relation actually exists in both languages for some percentages of occurrence for Process Planning.

In episode 5, the pattern differs from episode 4. The relation between the occurrence of Process Planning and text quality in episode 5 is positive in L1 and negative in L2, according to table 6. However, Process Planning occurs more, on average, in L2 (3.12%, see table 4) than in L1 (1.96 %, see table 4). So, if Process Planning activities were to occur in 3% of the segments in both L1 and L2, this would mean 'doing less of it than average' in L2, which is related to an increase of L2 text quality, while it would mean 'doing more of it than average' in L1, which is also related to an increase of text quality. Applying Process Planning activities in equal amounts during L1 and L2 writing actually has a similar effect on text quality in L1 and L2 if the amount lies between 1.96% and 3.12%. If, in both L1 and L2, Process Planning occurs in more than 3.12% or less than 1.96% of the segments in episode 5, its effect on text quality in L2 differs from its effect on text quality in L1. So, while table 6 seems to indicate that Process Planning is related to text quality differently in L1 and L2 (i.e. positively in L1 and negatively in L2) the relation is actually similar in L1 and L2 for some percentages of occurrence for Process Planning.

Comparing L1 and L2 directions of relations between text quality and Reading the Assignment, and Evaluating Own Text

For those cognitive activities which are applied similarly (i.e. averages per episode are not different) during L1 and L2 writing, interpreting the directions of relations is more straightforward. These activities are Reading the Assignment, Reading Own Text, Evaluating Own Text and Revising. As the average occurrence of these activities per episode is similar in L1 and L2, the L1 and L2 directions can be compared directly. As Reading Own Text and Revising are never related to text quality in the same episodes during L1 and L2 writing, a comparison of L1/L2 differences in directions is not made for these two activities, but only for Reading the Assignment and for Evaluating Own Text.

In episode 2, the occurrence of Reading the Assignment is positively related to text quality in L1, but negatively in L2. In this episode (actually, in each of the five episodes), Reading the Assignment is equally likely to occur in L1 and L2: in 5.33% of the segments. Applying these activities with a percentage higher (for example: 7%) than the mean occurrence in episode 2 implies 'doing more of it than average' in both L1 and L2. However, this leads to an increase in text quality in L1, but a decrease in text quality in L2. So, the positive relation in L1 and the negative relation in L2 remain, regardless of the percentage with which Reading the Assignment occurs. Reading the Assignment more than average in episode 2 is always related to an increase of text quality during L1 writing and related to a decrease of text quality during L2 writing.

The same line of reasoning applies to Evaluating Own Text in episodes 2, 4 and 5. In episodes 2, 4 and 5, the occurrence of Evaluating Own Text is positively related to text quality in L1, but negatively in L2. In these episodes, Evaluating Own Text is equally likely to occur in L1 and L2: in 0.61% (episode 2), 0.75% (episode 4) and 1.45% (episode 5) of the segments, respectively. Applying these activities with a percentage higher (for example: 2%) than the mean occurrence implies 'doing more of it than average' in both L1 and L2. However, this leads to an increase in text quality in L1, but a decrease in text quality in L2. So, the positive relation in L1 and the negative relation in L2 remain, regardless of the percentage with which Evaluating Own Text occurs. Evaluating Own Text more than average in episode 2 is always related to an increase of text quality during L1 writing and related to a decrease of text quality during L2 writing.

DISCUSSION

L2 text quality was found to be substantially lower than L1 text quality (*Cohen's d* > 1). The overlap in text scores (i.e. high scoring L2 texts and low scoring L1 texts) was relatively small, indicating that the L2 drop in text quality scores (relative to L1 text quality) holds for the majority of students in the population. The contribution of cognitive activities to text quality varies across the writing process in both L1 and L2, but the pattern of effectiveness of cognitive activities is generally different in L1 and L2 writing. Two differences are observed in the results:

1. Variations in the occurrence of cognitive activities are related to variations in text quality at different stages (i.e. episodes) of the writing process in L1 and L2 writing for all seven activities. For example: the occurrence of Content Planning is related to text quality at the start (episode 1) and during the middle part (episode 3) of the writing process in L1, whereas in L2, the occurrence of Content Planning is related to text quality in later stages of task execution (episode 4). This means that, for L1 and L2 writing, there are different crucial moments during task execution, at which it matters whether writers apply a specific cognitive activity more or less often. In other words, writers need to distribute their attention differently across L1 and L2 writing tasks.
2. Where variations in the occurrence of a cognitive activity are related to variations in text quality in the same episode(s) during L1 and L2 writing, these relations are most often positive in L1, but negative in L2. Such dissimilar L1 and L2 relations hold for Reading the Assignment in episode 2, and for Evaluating Own Text in episodes 2, 4 and 5. For Process Planning, it sometimes holds true in episodes 4 and 5, depending on the amount of Process Planning applied. In episode 4, L1 and L2 relations are dissimilar (i.e. positive vs. negative) if Process Planning occurs in less than 1.44% or more than 1.29% of the segments. In episode 5, L1 and L2 relations are dissimilar if Process Planning occurs in more than 3.12% or less than 1.96% of the segments. Only in a minority of cases, the effect of an activity is similar for L1 and L2 writing: in episode 4, if Process Planning occurs in more than 1.44% or less than 1.29% of the segments and in episode 5 if Process Planning occurs in less than 3.12% or more than 1.96% of the segments.

Returning to the main question of the present study (research question 2), the obtained results show that, in general, the contribution of cognitive activities to text quality is distributed differently across task execution in L1 and L2 writing (table 6). This means that, for the age group under study, effective L2 writing processes are different from effective L1 writing processes. L2 text quality is in general lower if writers stick to their L1 orchestration of cognitive activities during L2 writing tasks.

The finding that cognitive activities are related to text quality at different stages (i.e. episodes) of the writing process in L1 and L2 is in line with Van Weijen's (2009) results. Although the L1 and L2 relations between writing processes and text quality were not directly comparable in Van Weijen's (2009) study, she also found that relations (reflected by negative or positive correlations) between cognitive activities were stronger at some stages of the writing process than at others, and that the stages at which a strong correlation exists are not always similar for L1 and L2 writing. However, there are also differences between Van Weijen's (2009) findings and the results of the present study. Reading the Assignment, for example, is negatively related to text quality at the start of L2 task execution in the present study, but positively correlated to text quality at the start of the L2 writing process in Van Weijen's (2009) study. Such differences are possibly explained by the different populations under investigation: first-year university students in Van Weijen's (2009) study, and students of secondary education in the present study.

The finding that cognitive activities are regularly positively related to text quality in L1, but negatively in L2 is an observation which raises new questions. It was established that applying a cognitive activity in equal amounts during L1 and L2 writing can, in specific episodes, be related to an increase of text quality in L1, but to a decrease of text quality in L2. (And indeed, all-but-one of the established relations between cognitive activities and text quality in L2 are negative relations.) How can it be the case that something which is a feature of successful L1 writing (e.g. evaluating your own text in episodes 2, 4 and 5), becomes a feature of less proficient writing in L2? At this point, we can only speculate about an explanation. The most notable difference between the L1 and L2 conditions is that during L2 writing, students' language proficiency is generally of a lower level than during L1 writing. Although students focused on cognitive activities during the same stages of L2 task execution, and in equal amounts, as during L1 task execution (table 4), it is possible that, due to lower second language proficiency, the cognitive activities which are applied during L2 writing are lacking on different aspects of quality. In other words, students' L1 and L2 writing processes are of equal quality in terms of temporal organizations (and amount of segments – or attention - applied to evaluating). Nevertheless, it is possible that temporal distributions reflect the quality of processing to a large extent, but not completely, as seems to be suggested by the moderate correlations between predicted and observed values in table 4. There might, in short, be additional features of processing quality on which L1 and L2 writing processes can differ. Two such features, which are partly interrelated, are discussed below.

First, the quality with which cognitive activities are executed may be expressed in terms of their objects. Reading the formal goals specified in an assignment (e.g. length and audience of the intended essay) is qualitatively different from reading additional documentation on the topic. Similarly, Evaluations of Own Text can have various objects. Evaluating local text features (e.g. spelling) is a qualitatively different activity from evaluating global text features (e.g. coherence of the presented argumentation), for example. If L1 and L2 Evaluations of Own Text differ in terms of their objects (for example, if global evaluations are neglected in L2), this could explain why they are positively related to text quality in L1, but negatively in L2.

Second, the quality with which cognitive activities are executed may be expressed in terms of their interrelations with other cognitive activities, which precede or follow it. Van den Bergh and Rijlaarsdam (1999) showed that the function of a cognitive activity can be expressed by investigating what kind of activity precedes it. They distinguish different kinds of 'content generating', such as 'Assignment-Driven-Generation' (new ideas are sparked by reading information in the assignment) and 'Translation-Driven-Generation' (new ideas are sparked as the writer formulates text; the activity of putting thoughts into language can be an incentive for generating new content). In the same vein, the function and quality of Reading the Assignment might be defined by the activity following it. Reading the Assignment to generate ideas is a qualitatively different activity than Reading the Assignment to understand the goals of the assignment. Indeed, if Reading the Assignment with the goal of understanding the assignment is performed in later stages of the writing process, it is likely to be a symptom of a problematic writing process. Possibly, Reading the Assignment in episode 2 is more often reading-to-understand-goals in L2 (hence, possibly, the negative relation between its occurrence and text quality in episode 2 during L2 writing), and more often reading-to-generate in L1 (hence, possibly, the positive relation between its occurrence and text quality in episode 2 during L1 writing). Similar lines of reasoning can be applied for other cognitive activities. The quality of Evaluating Own Text, for example, might be characterized by whether or not it is followed by a (successful) revision.

The explanations provided above could be incorporated in future research. In the first place, the interrelations between cognitive activities and the objects of cognitive activities could be incorporated in operationalizations of the quality of writing process execution, in addition to a temporal approach. Furthermore, the influence of language proficiency could be investigated. If it is indeed the case that language proficiency affects the quality with which cognitive activities are executed, then it is to be expected that, for higher L2 proficiency scores, the L2 relations between cognitive activities and text quality are more similar to their L1 counterparts.

The study presented here unites three methodological advantages which had, to our knowledge, not yet been combined in a study on L1/L2 writing. First, multiple tasks are used per language, so that we can separate language effects from task effects. Second, the writing processes, and relations between writing processes and text quality, are analyzed in terms of the temporal distributions of cognitive activities, so that the quality of writing processes was operationalized in a sensitive and valid manner. Third, L1 and L2 text quality are expressed on a single scale. This allowed us to make a direct comparison of the L1 and L2 relations between the writing process and text quality: do the underlying writing processes contribute to text quality differently during L1 and L2 writing? This was established to be the case, which means that L2 text quality is in general lower if writers stick to their L1 orchestration of cognitive activities during L2 writing tasks.

CHAPTER 4

THE EFFECT OF LINGUISTIC PROFICIENCY ON L1 AND L2 WRITING PROCESSES

Abstract

Language proficiency is, in most theoretical writing process models, considered a constituent part of writing proficiency. Particularly for L2 writing, language proficiency has been regarded as an important explanatory variable. Studies in which writing process differences are explained by individual language proficiency measures in L1 and L2 are scarce. The present study sets out to do this, by explaining writing process differences in L1 and L2 with individual L1 and L2 language proficiency scores. Twenty fourteen- and fifteen-year-olds wrote three or four essays in L1 (Dutch) and three or four essays in L2 (English, taught as a foreign language). They also completed an L1 and an L2 language proficiency test. Students' writing processes were registered by means of think aloud procedures, combined with keystroke logging. Writing processes were analyzed temporally: the moment at which cognitive activities (such as planning, formulating, and revising) occur during writing (i.e. at the start or end of task execution) reflects the quality of the writing process. Results indicate that language proficiency had an effect on the occurrence of Evaluating Own Text in L1, and on the temporal distribution of Process Planning in L2.

A writing process can be characterized by a specific configuration of the constituent cognitive activities, such as generating content, planning, formulating (text production) and making revisions when needed, while keeping rhetorical goals in mind and staying aware of the intended audience (Flower & Hayes, 1980; Hayes & Flower, 1980; Hayes, 1996). The writing process is not a linear chain of actions in which planning, generating, text production and revising, for example, are carried out in consecutive phases. Rather, all these cognitive activities can (re)occur at (more or less) any given moment during the writing process.

Indeed, it has been shown that the quality of writing process execution is reflected by the temporal distribution of cognitive activities across task execution: it matters at which moment during the writing process specific cognitive activities occur (Breetvelt et al., 1994; Leijten & Van Waes, 2006; Olive et al., 2008; Roca de Larios, Marín & Murphy, 2001; Rijlaarsdam & Van den Bergh, 1996; Rijlaarsdam et al., 2011; Van den Bergh & Rijlaarsdam, 1996; Van den Bergh et al., 2009; Van der Hoeven, 1997). Rijlaarsdam and Van den Bergh (1996), for example, demonstrated that cognitive activities' contribution to text quality varies across the writing process. For example, structuring activities were more effective when they occurred during early stages of task execution (i.e. the correlation between structuring and text quality is at its highest at the start of the writing process) and less effective when they occurred towards the end of task execution (Rijlaarsdam & Van den Bergh, 1996). In other words, writing processes are more effective if structuring activities are applied predominantly at the start of task execution (and less effective if this is not the case). Differences in text quality, then, are related to differences between writers in terms of the degree to which they adhere to the more effective temporal distribution of cognitive activities (such as structuring) across the writing process.

Van Weijen (2009, p. 96), who investigated the writing behavior of first-year university students, showed that effective distributions of cognitive activities for first language (L1) writing are different from effective distributions for second language (L2) writing. In L1, for example, the correlation between generating and text quality is highest at the start and lowest at the end of the writing process, whereas in L2, this correlation is more or less constant across the entire writing process. There can be individual variations: for individual writers, the most effective distribution can be different from the distribution which is the most effective on average. Nevertheless, Van Weijen (2009, p. 96) showed that the distribution which is the most effective on average, is effective for the majority of students. In L1, for example, it holds for (at least) eighty percent of the population that generating activities should be applied more at the start and less at the end of task execution.

In both L1 and L2, though, different writers show different temporal distributions (Van Weijen, 2009, p. 92-93), indicating that the quality of L1 and L2 writing processes differs. In the L1, for example, most students do indeed apply the majority of their generating activities at the start of task execution, but some do not: they apply structuring activities during middle parts of the writing process, or even towards the end. Van Weijen (2009) showed that this variation between writers, in terms of temporal distributions, exists in both L1 and L2, but that inter-individual differences are larger during L2 writing than during L1 writing.

An important question is: what explains the fact that writing processes differ between writers? After all, the answer to this question would enhance our understanding of the constituent factors (knowledge, skills, et cetera) underlying writing proficiency. Of course, there are multiple possible explanations available for process variation within writers. Inter-individual variation can, for instance, be induced by differences between tasks (Rijlaarsdam et al., 2011; Van den Bergh et al., 2009; Van Weijen et al., 2008). Van Weijen et al. (2008), for example, demonstrated that tasks which differ only in terms of topic cause significantly different distributions of cognitive activities. Nevertheless, in Van Weijen et al.'s (2008) study, variation between individual writers was larger than variation between tasks (within individuals). This suggests that characteristics of individual writers are relevant for explaining differences between writing processes.

Hayes' (1996) theoretical model for describing the writing process identifies a number of resources which the individual should possess and which are assumed to affect the way in which the writing process is executed. Three main resource components within the individual, according to this model, influence the writing process: motivation, working memory, and long-term memory. The latter comprises knowledge elements, such as topic knowledge, genre knowledge and linguistic knowledge.

A number of researchers have been interested in the role of linguistic proficiency. Van der Hoeven (1997, p. 112-115), for example, showed that linguistic skill (measured by a test comprising items on lexical knowledge, reading comprehension and sentence production) explains some of the differences between 12-year-old students' distributions of cognitive activities during L1 writing. At the start of task execution, students with higher linguistics skill were found to generate more than average, while those with lower linguistic skill generated less than average. Across the writing process, these differences between students with higher and lower linguistic skill grow smaller. Formulating activities are applied less than average at the start of task execution by students with higher linguistic skill, but more than average at the end of task execution (and vice versa for students with lower linguistic skill). Structuring activities were applied more by students with higher linguistic skill than by students with lower linguistic skill, and this effect seemed to be constant across the whole writing process. No effect of linguistic skill was found on reading the assignment, monitoring, metacommenting, re-reading own text, evaluating and revising. These findings seem to illustrate that linguistic skill does not only have an impact on those cognitive activities which are largely language-specific, such as formulating, but also on cognitive activities which are less language-specific, such as structuring.

While Van der Hoeven's (1997) results only apply to L1 writing, language proficiency has been regarded as a particularly important explanatory variable for inter-individual differences in L2 writing (Chenoweth & Hayes, 2001; Sasaki & Hirose, 1996; Schoonen et al., 2003). After all, differences between students' language proficiency level are often larger in L2 than in L1. In fact, explanations for the often observed lower quality of L2 writing (as compared to L1 writing) often revolve around language difficulties (Chenoweth & Hayes, 2001; Sasaki & Hirose, 1996; Schoonen et al., 2003). Students' lower language proficiency in L2 is, in such explanations, assumed to limit their ability to express their ideas. In addition, language difficulties constrict working memory resources, leaving fewer resources for conceptual and regulatory activities (such as structuring and monitoring) (cf. McCutchen, 1996, who describes this mechanism for L1 writing). The latter mechanism has been forwarded as a 'threshold hypothesis' (Sasaki & Hirose, 1996; Schoonen et al., 2003): a certain level of L2 language proficiency should be attained before students are able to apply conceptual and regulatory activities, or to transfer their L1 writing strategies to L2 writing situations.

Schoonen et al. (2003) compared the contribution of language proficiency (linguistic knowledge: vocabulary, grammar, orthography, and linguistic speed: lexical retrieval and sentence building speed) to the quality of writing between L1 and L2 writing among 13- and 14-year-old students. They conclude that L2 linguistic proficiency explains more of L2 text quality than L1 linguistic proficiency explains of L1 text quality. This suggests that language proficiency plays a larger role in L2 writing than in L1 writing. However, whether the quality of processing was affected by lower language proficiency could only be inferred, as no process data were available in Schoonen et al.'s (2003) study. This is also the case for Sasaki and Hirose's (1996) study. They investigated the influence of language proficiency on L1 and L2 text quality, but not on L1 and L2 writing processes.

There are a number of studies in which L1 and L2 writing processes are compared (Chenoweth & Hayes, 2001; Hirose, 2003; Lindgren, Spelman Miller & Sullivan, 2008; Van Weijen, 2009), but in none of these studies any data are available on general language proficiency. It is often assumed that the L2 condition reflects lower language proficiency levels. This is usually true, but any differences between L1 and L2 writing processes cannot automatically be ascribed to language proficiency differences, since L1 and L2 writing situations vary on other aspects too. For example, writers may also have less L2 genre knowledge, or less knowledge of the L2 culture. Roca de Larios et al. (2001) compared L1 and L2 writing processes of students at three different levels in the education system. While these three groups are likely to vary in terms of language proficiency, they are, however, likely to vary on other aspects, too, such as world knowledge and audience awareness. So again, differences between writing processes could not be solely attributed to language proficiency differences.

In short, studies in which individual writing process differences are explained by individual language proficiency measures in L1 and L2 are scarce. The present study aims to gain more insight into this issue, by explaining writing process differences in L1 and L2, in terms of temporal distributions, with individual L1 and L2 language proficiency scores. It is expected that language proficiency is a constituent part of both L1 and L2 writing, and has an effect on both language-specific and non-language-specific cognitive activities.

METHOD

Participants

The participants were fourteen- and fifteen-year-old students (N = 20; 10 female and 10 male). They were from three different third-year-forms at the same school for pre-university secondary education. They were recruited by means of a call for volunteers, which was distributed by their Dutch language teacher. All participants were native speakers of Dutch. On average, students had followed English as a school subject for four or five years, for approximately two or three hours a week. Participants received a financial reward for their participation. Parental consent was obtained.

Instruments and procedures

Writing tasks

The students completed eight writing tasks. They wrote four argumentative essays in L1 (Dutch), and four argumentative essays in L2 (English), on topics such as 'camera surveillance in inner city areas' or 'downloading music'. Multiple tasks were used per language, in order to be able to disentangle task effects and language effects. After all, if only one task were used per language, it would be impossible to know if any differences which are found are due to task or due to language (Van den Bergh et al., 2009; Van Weijen, 2009).

All eight writing assignment were similar in terms of audience (peers), medium (a school-related magazine for secondary school students) and purpose (to convince the readers of your point of view), and differed only in terms of topic. The essays had to be about half a page (A4 format) in length (which is about 250 to 300 words). The assignments were tested with third year students of pre-university secondary education during a pilot study in 2005. They were also successfully used in previous studies (Van Weijen et al., 2008; Van Weijen, 2009). An example of an assignment can be found in Appendix A.

The available time for each essay was approximately thirty minutes, although participants were allowed to go on longer if they felt that their essays were not finished yet. No participant used more than forty minutes. The students completed the essays during two separate days. Between the four tasks completed per day, participants were given a short break of about ten to fifteen minutes.

To avoid sequence effects and to control for effects of topic, several measures were taken. It has been established that topic can greatly influence the writing process (Rijlaarsdam et al., 2011; Van den Bergh et al., 2009; Van Weijen, 2009). Therefore, to disentangle language effects from topic effects, the topics were systematically balanced across languages, so that each topic occurred in both L1 and L2. So, writer 1 wrote on topics 1, 2, 3 and 4 in L1 and on topics 5, 6, 7 and 8 in L2, whereas writer 2 wrote on topics 2, 3, 4 and 5 in L1 and on topics 1, 6, 7, and 8 in L2, and so forth. The order in which L1 and L2 essays were written was also balanced across participants: ten students first completed four L1 essays, and then wrote four L2 essays; the other ten students first completed the L2 essays, and then wrote the L1 essays.

The students wrote the essays on a computer using Microsoft Word, in the presence of a test leader. They all were very familiar with using MS Word. They had to think aloud during the process of task execution. If they fell silent, the test leader prompted them to continue thinking aloud by a neutral remark: "aloud, please" (in Dutch). The participants practiced thinking aloud before writing their essays by means of a short mathematical puzzle and a two-line writing assignment. All writing sessions were audio- and video-taped and recorded by means of keystroke logging (Inputlog: Leijten & Van Waes, 2006).

Coding process data

All think aloud data were transcribed and segmented, and completed by the Inputlog recordings. A new segment in the protocols reflected a switch to a different cognitive activity within a participant's writing session (Van den Bergh & Rijlaarsdam, 2001). All segments were coded according to a coding scheme (adapted from Breetvelt et al., 1994). The coding scheme (see table 1) consisted of eight main categories. 'Revising', as a category, involved subcodings. Revisions were subcoded as 'automated corrections' when they involved corrections of typographic errors. They were typically errors which are made as a result of the use of a keyboard, which seem to be corrected almost automatically. An example would be: a writer types 'almots', and immediately corrects this error by giving to backspaces and typing 'st', so that it now reads 'almost'. All revisions which were not 'automated corrections' were subcoded as 'conceptual revisions': they involve alterations which are made by the writer in a non-automated way. That is, they involve actual alterations at the level of spelling or content. In the remainder of this article, we will use the term 'revision' or 'revising' only if we are referring to conceptual revisions. One randomly chosen think aloud protocol (number of segments = 518) was coded by two researchers. The intercoder agreement (Kappa) was satisfactory (.85).

Table 1.Coding categories in the coding scheme

	Coding category	**Description**	**Example**
READING THE ASSIGNMENT	Reading the instruction text and documentation	Reading (part of) the task instructions	"Write an essay in which you..."
PROCESS PLANNING	Monitoring	Verbalizations indicating a steering capacity which governs the writing process, mostly self-instructions	"I'm going to read what I've written so far."
	Metacomments	Evaluations of a student's own writing process	"I should have made an outline of the text before I started."
CONTENT PLANNING	Goal setting	The formulation of goals which the text has to satisfy	"The text should be convincing."
	Generating	Generating ideas for content or form	"Something about the disadvantages of camera surveillance..."
	Structuring	Evaluating and arranging ideas	"Something about adults? No, that's not relevant."

Table 1 (continued). Coding categories in the coding scheme

FORMULATING	Text production.	Production of new text	"Camera surveillance invades people's privacy." (Verbalizations usually occur parallel to the activity of typing.)
READING OWN TEXT	Reading produced text	Reading (part of) the produced text, at any given moment during the writing process	" Surveillance cameras do not increase public security."
EVALUATING OWN TEXT	Evaluating produced text	Evaluation of produced text	"The largest part is about backdraws."
REVISING	Revising	Making changes (at word, sentence or text level) to the text produced so far	Moving a set of sentences from the body of the text to the introduction.
	Automated corrections	Corrections of typographic errors due to keyboard use. *(Not included in the analyses.)*	Writer types 'almots', deletes 'ts' by means of 'backspace', en then types 'st'. (Mostly no (explicit) verbalization of the correction.)
OTHER	Pauses	Silence or interjection	"eeeerrr"
	Interaction with test leader	Interaction between test taker and test leader	"Could I open a window?"
	Physical activity	Physical activity	Taking a sip of tea.
	Navigation	Moving through the document: arrow buttons or mouse movements	Moving the cursor some lines back.

As we are interested in complete writing processes, all cognitive activities in table 1 are analyzed, except for the activities listed under 'Other'. The reason for excluding the 'Other' activities is that they are either too diffuse (Pausing and

Navigating can reflect all kinds of cognitive processing, or none at all) or conceptually irrelevant and highly infrequent (Interaction With Test Leader and Physical Activity). Of the seven remaining categories, three are pre-eminently language-specific, namely Reading the Assignment, Formulating and Reading Own Text. Process Planning is a largely non-language-specific activity. The other three activities under analysis (Content Planning, Evaluating Own Text and Revising) combine language-specific and non-language-specific elements.

Table 2 shows part of a protocol and illustrates how the think aloud data en the Inputlog data were integrated. The protocols consisted of seven columns. The column labeled 'Reading' was used for indicating if any reading activities (reading the assignment - RA – or reading own text – ROT) were taking place. The 'Verbalizations' column contained everything which was said out loud by the student, except interjections, such as "uhm". All information in the column labelled 'Typing' are derived from Inputlog. There were three categories in this column, namely the production of new text, revisions (indicated by Inputlog as [BS] for backspace or [DEL] for if the delete button was pressed), and navigation (by means of mouse movements or arrow buttons). The 'Pausing' column contained all silences and interjections. The column labeled 'Other' mostly contained descriptions of physical activities, for example 'takes a sip of his drink'. One row is one protocol segment. As such, this transcription method allows for parallel actions. Text production and verbalizations, for example, often occur simultaneously. The 'Typing' column would contain the text production as registered by Inputlog. The codes in the last column were in reality numbers. Code

Table 2. Part of a completed protocol

Segment number	**Reading**	**Verbalizations**	**Typing**	**Pausing**	**Other**	**Code**
17		even een titel erboven *(I'll insert a title)*				Monitor
18		having	Having childere			Text production
19		chil	[BS 1] [BS 1] [BS 1]			revision (automated correction)
20		children yes or no	ren, yes or no?			text production
21	ROT	having children yes or no				Reading own text

01, would stand for 'reading the assignment', for example, and code 02 would stand for monitoring. English translations of Dutch verbalizations are given in italics.

Due to technical deficits (defective videotapes or unrecorded keystroke loggings), sometimes not all eight writing tasks administered per participant were available for analysis. Table 3 shows the exact number of tasks included in the analysis per language condition and per participant. It is clear that, while the measurement sometimes involves less than four tasks per language condition per student, the writing processes of all students have been measured using multiple tasks per language, which enhances the measurement's reliability.

Table 3. Number of tasks included in the analysis

Participants	**L1**	**L2**
2, 3, 4, 5, 7, 8, 10, 11, 13, 15,	4	4
9, 12, 14, 16, 17, 18, 20	4	3
6, 19	3	3
1	2	4

Measuring language proficiency

Students' general language proficiency was approximated by administering vocabulary tests, in both L1 and L2. Particularly for L2 situations, vocabulary size is sometimes seen as one of the best predictors of language proficiency (Laufer & Goldstein, 2004). For both L1 and the L2, the vocabulary tests consisted of sentences with blanks, which participants had to fill in, using one word per blank. In the L1 tests, the looked-for word could be inferred from the sentence. Table 4 shows examples of items from both the L1 and the L2 language proficiency tests. In the L2 tests, its Dutch translation was given between brackets. The complete tests can be found in Appendix G.

Table 4. Examples of items in the language proficiency tests.
For the Dutch L1 item, the English translation is provided in italics

	Item	**Solution**
L1	... hij een druk leven leidt, maakt hij altijd tijd voor me vrij. *… he has a busy life, he always makes time for me.*	Hoewel *Although*
L2	I have to practise the piano every day ... [of] I like it or not.	whether

On the basis of a pretest, the items in the test were ordered from easy to difficult. They were timed tests: the students were given six minutes to complete a 62- (for

L1) or 64-item (for L2) test. Internal consistency of both tests was high (α = .94 for L1 and α = .85 for L2).

Analyses

Analyses of writing processes should take 'time during the writing process' into account as an explanatory variable to operationalize the writing process in a valid and sensitive way (Rijlaarsdam & Van den Bergh, 1996; Van den Bergh & Rijlaarsdam, 1996). In the present study, this was accomplished by splitting each protocol into five equally long episodes in terms of numbers of segments (cf. Breetvelt et al., 1994; Roca de Larios et al., 2008; Van den Bergh & Rijlaarsdam, 1996). A protocol of 330 segments, for example, would be analyzed as five episodes consisting of 66 segments each. For each cognitive activity under analysis (i.e. reading the assignment, process planning, content planning, text production, reading own text, evaluating own text, and revising) its proportion of occurrence relative to the total number of segments was calculated for each episode. For instance, if an episode consisted of 80 segments (which means that the entire writing process consisted of 400 segments), and 10 of these segments were coded as 'reading the assignment', then the proportion for reading the assignment in that episode would be 0.125. This way, the frequency with which a cognitive activity occurs is allowed to vary across the writing process.

A multilevel regression model has been applied to model the occurrence of the cognitive activities at each episode, as episodes are nested within writers and tasks (Van den Bergh et al., 2009). In effect, a longitudinal model is in operation, as it concerns changes in occurrence during the writing process: proportions of the applied cognitive activity may be different during each new episode. Therefore, the occurrence of each of the seven cognitive activities (A) had to be described as a function of episode, i.e. A=*f*(episode). Note, however, that this function *f* does not need to be identical for all individuals *i* and tasks *j*: $A=f_{ij}$(episode).

This function, *f*, can take many forms (Goldstein, 1979; Healy, 1989). For this study, polynomial models were preferred because of their flexibility. Depending on the number of coefficients (and their numerical values), polynomials can take almost any shape. As such, they can be used to model various kinds of growth patterns.

Growth across task execution is not necessarily linear. For instance, text production activities may occur relatively little during first and last episodes of the writing process, but a lot during the middle part of task execution (e.g. during episode 3) Therefore, non-linear terms (e.g. quadratic or cubic terms) can also be included in the model: the occurrence of an activity (at each episode) is described as powers of episode (episode^0, episode^1, episode^2, …). The number of parameters needed to describe the observed activities (in each episode) is considered an empirical matter. That is, a next power of episode is included in the model only if it has a significant contribution in the description of an activity *and* if all lower powers are significant as well (see, Van den Bergh & Rijlaarsdam, 1996). For example, 'episode' to the second power can only be added to the model, if the linear term (episode^1) has reached significance.

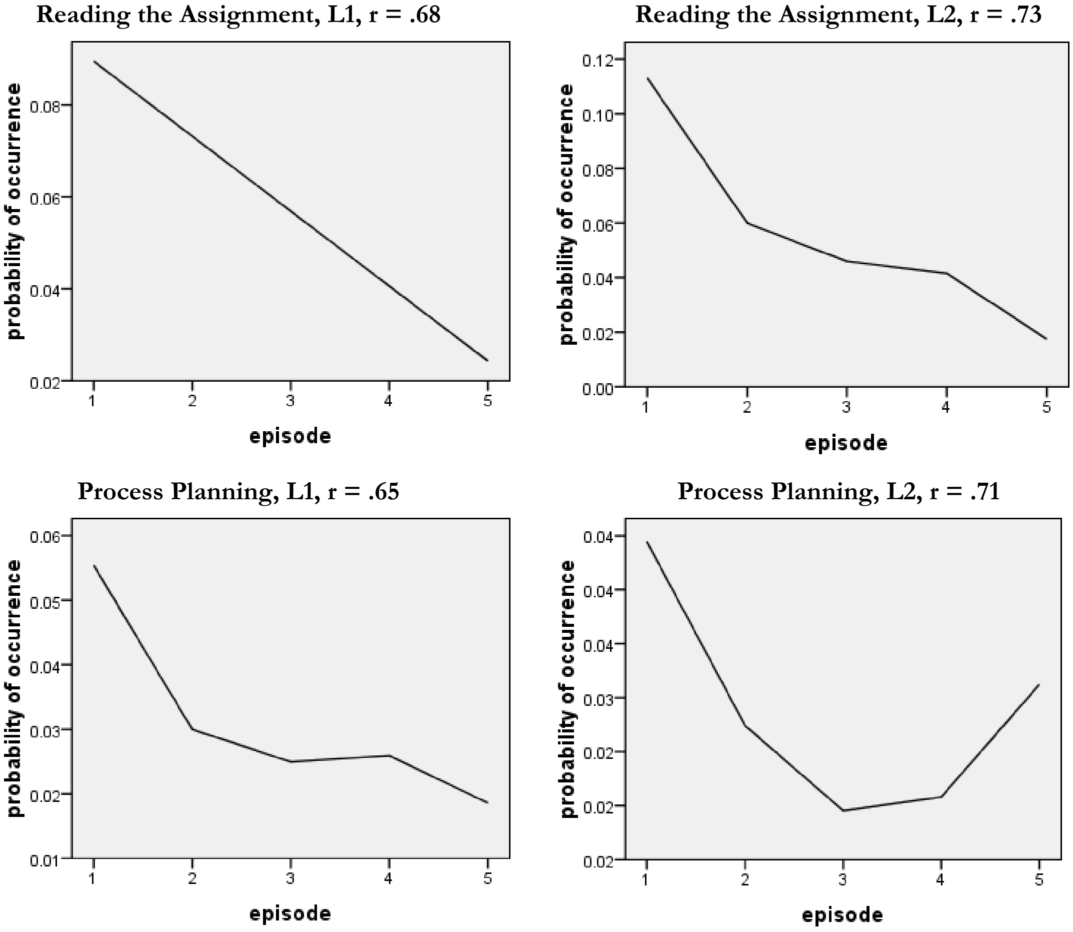
To meet the requirement that the function *f* is allowed to differ between individuals, not only are the regression coefficients of the powers of episode estimated, but also the variance of these parameters. That is, the variance of the intercept (writers differ in occurrence at episode = 0) and the variance of the linear component (writers differ in linear change over the writing process). These variance components are in fact the variances of residuals which characterize the occurrence of activities of a specific writer. The differences between individuals can be explained by individual characteristics like their language proficiency scores. Adding these scores, then, is the final step in the construction of the regression model. In this model, then, episode and the individual language proficiency scores (as z-scores) were the explanatory variables. Of course, the effect of the language proficiency scores is not (necessarily) constant across task execution. (For instance: differences between students in terms of the proportion with which Reading the Assignment occurs, as explained by language proficiency scores, might be expected to be larger at the start of task execution than during later parts.) Therefore, interaction effects between the language proficiency scores and the time variable (episode) were also calculated. The complete multilevel regression model, as used for explaining the occurrence of each of the seven cognitive activities, can be found in Appendix H. The regression coefficients were estimated for L1 and L2 separately. When explaining L1 writing processes, the L1 language proficiency scores were used in the analysis, and when explaining L2 writing processes, L2 language proficiency scores were used.

RESULTS

Figure 1 presents the average distributions of each of the seven cognitive activities across task execution, for both L1 and L2 writing tasks. The regression weights can be found in Appendix I. To assess the degree to which the regression models (with predictor variables episode^0, episode^1, episode^2, and so forth) explain the total variance in the process data, we correlated the values as predicted by the models and the observed values. These correlations (*r*) are also presented in figure 1. All correlations are satisfactory, ranging from .59 to .87. This means that 'time during the writing process' explains a substantial part of the variation in the data.

The occurrence of Reading the Assignment, Process Planning, Content Planning, Formulating and Evaluating Own Text varies across task execution in both L1 and L2. Reading the Assignment is most likely to occur at the start of the writing process and least likely to occur at the end of the writing process, both in

Figure 1. Average distributions of cognitive activities across task execution. r = *correlation between values as predicted by the models and values as observed*



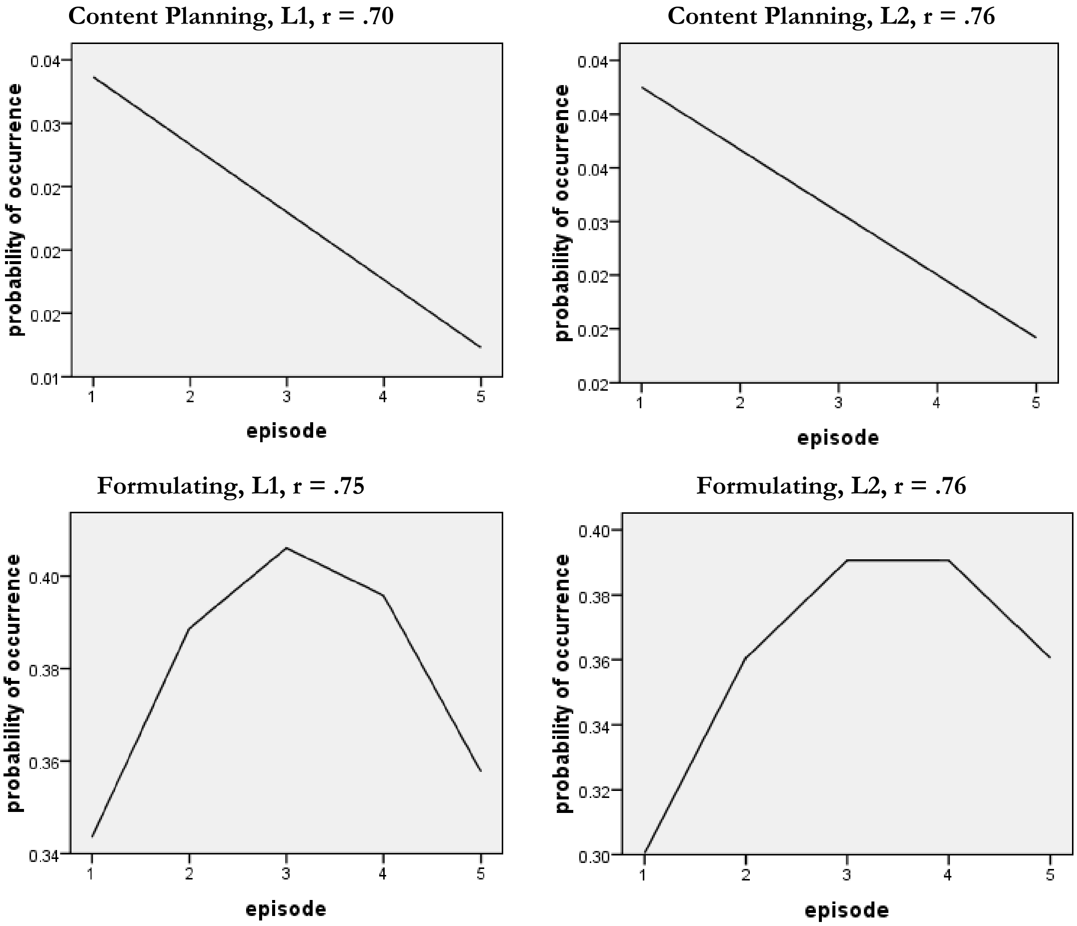
Reading the Assignment, L1, r = .68

Reading the Assignment, L2, r = .73

Process Planning, L1, r = .65

Process Planning, L2, r = .71

Figure 1 (continued). Average distributions of cognitive activities across task execution. r = *correlation between values as predicted by the models and values as observed*



Content Planning, L1, r = .70

Content Planning, L2, r = .76

Formulating, L1, r = .75

Formulating, L2, r = .76

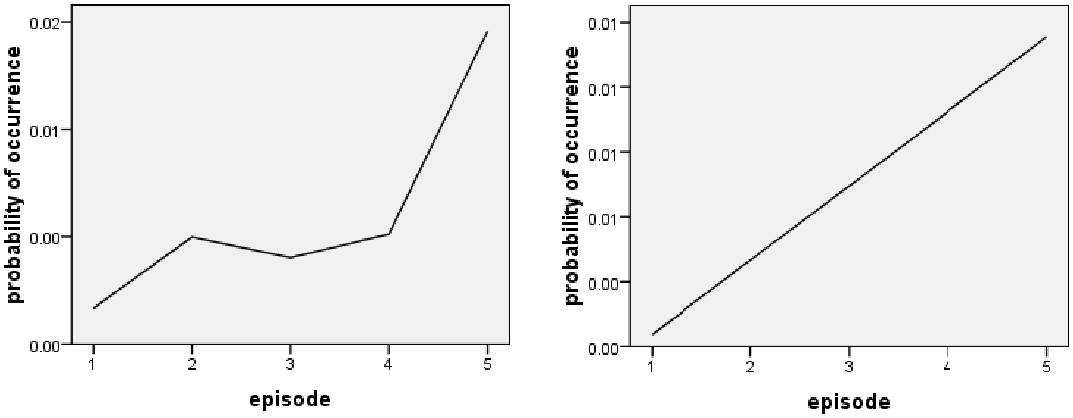
Reading Own Text, L1, r = .87

No effect of episode.
Average probability of occurrence = .02

Reading Own Text, L2, r = .78

No effect of episode.
Average probability of occurrence = .01

Figure 1 (continued). Average distributions of cognitive activities across task execution. r = *correlation between values as predicted by the models and values as observed*



Evaluating Own Text, L1, r = .62

Evaluating Own Text, L2, r = .59

Revising, L1, r = .67

No effect of episode.
Average probability of occurrence = .05

Revising, L2, r = .69

No effect of episode.
Average probability of occurrence = .04

L1 and L2. Process Planning is also most likely to occur at the start of the writing process. Its likelihood of occurrence decreases through the middle of the writing process (episode 3), then slightly increases (episode 4). In L1 writing, we then see a decrease towards the end of task execution (episode 5). In L2, on the other hand, the likelihood that Process Planning occurs increases at the end of task execution. Content Planning is, both in L1 and L2, most likely to occur at the start and least likely to occur at the end of the writing process. The distribution of Formulating activities is very similar for L1 and L2 writing during episodes 1 through 3. At the end of the writing process (episodes 4 and 5) the probability that Formulating occurs decreases more strongly in L1. Evaluating Own Text is least likely to occur at the start and most likely to occur at the end of the writing process, in both L1 and L2. In L1, a slight decrease of Evaluating Own Text was established during the middle part of writing. This could not be established for Evaluating own Text during L2 writing. For Reading Own Text and for Revising, no effect of different episodes was found: no variation in their occurrence across the writing process could be established in either L1 or L2. This means that the occurrence of these two activities is best described by their mean occurrence during the entire writing process.

The main question of the present study, however, was whether distributions of cognitive activities vary between individual writers according to their language proficiency scores. Table 5 shows descriptive information about the language proficiency tests. In both languages, there is sufficient spread of test scores across participants.

Table 5. Descriptive statistics for the L1 and L2 language proficiency tests: means (M), standard deviations (SD), minimum and maximum scores

	L1	**L2**
M	43.5	23.9
SD	6.5	5.3
Min.	31	14
Max.	57	34

Table 6 shows whether an effect of language proficiency on the occurrence of cognitive activities was found. An effect of language proficiency on the writing process was found for Evaluating Own Text in L1, and for Process Planning and

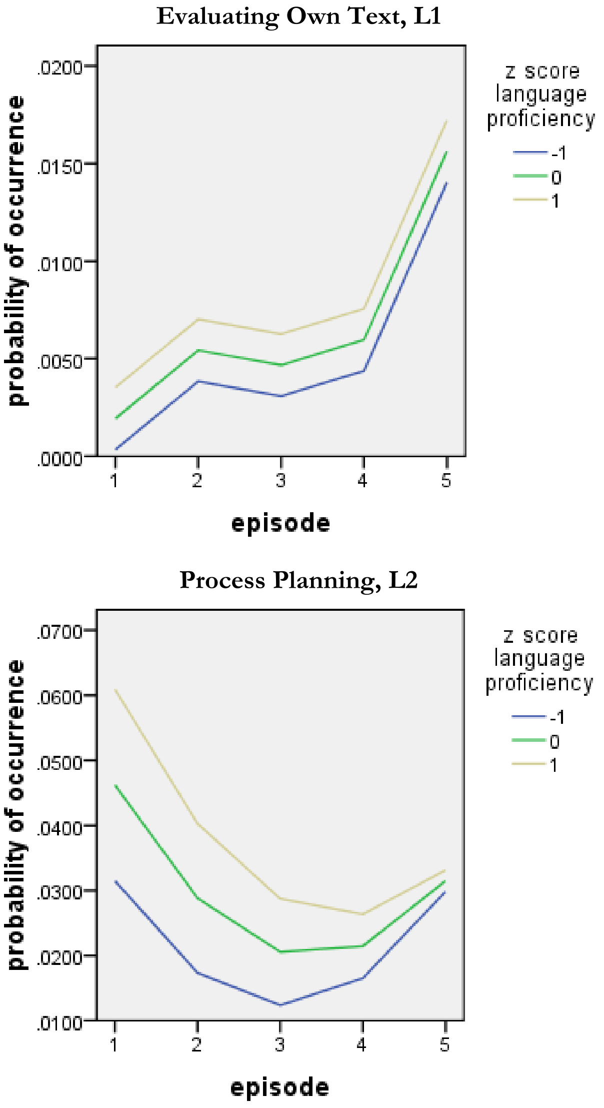
Table 6. Effects of language proficiency (LP) scores on cognitive activities during writing. V: a significant effect was established; empty cell: no significant effect was established

	L1		**L2**	
	Main effect: LP	*Interaction effect: LP*episode*	*Main effect: LP*	*Interaction effect: LP*episode*
Reading Assignment				
Process Planning			V	V
Content Planning				V
Formulating				
Reading Own Text				
Evaluating Own Text	V	V		
Revising				

Content Planning in L2. The regression weights can be found in Appendix J. For these three cases, the fit of the regression models with episodes and language proficiency as predictors and the regression models with only episodes was assessed by comparing the respective -2 log likelihoods. This procedure tells us whether language proficiency makes a unique contribution to the explanation of variation between writing processes. For Process Planning in L2, the model with episodes, language proficiency and language proficiency*episode as predictor variables is a better fit to the data than the model with only episodes ($\chi^2 = 6.46$, $df = 2$, $p < .05$). For Content Planning in L2, and for Evaluating Own Text in L1, the fit of the model does not improve by adding language proficiency and language proficiency*episode as predictors ($\chi^2 < 4.15$, $df = 2$, $p > .12$). However, for Evaluating Own Text in L1, a model with episode and language proficiency (but not the interaction variable language proficiency*episode) is a better fit than an episodes-only model ($\chi^2 = 6.3$, $df = 1$, $p < .05$). In short, language proficiency only explains variations in distributions of Process Planning in L2. In addition, language proficiency explains differences in the average occurrence of Evaluating Own Text. For all other cognitive activities, language proficiency makes no (unique) contribution to explaining inter-individual writing process differences.

Figure 2 illustrates the effect of language proficiency on L1 Evaluating Own Text and L2 Process Planning throughout the writing process. The middle curves reflect the (estimated) distributions of Evaluating Own Text or Process Planning for students with average L1 (Evaluating Own Text) or L2 (Process Planning) language proficiency test scores (*z score language proficiency = 0*). The upper lines reflect the distributions of Evaluating Own Text or Process Planning for students with language proficiency scores of one standard deviation above the average score (*z score language proficiency = 1*). The lower lines reflect the distributions of Evaluating Own Text or Process Planning for students with language proficiency scores of one standard deviation below the average score (*z score language proficiency = -1*). During L1 writing, students with higher (L1) language proficiency scores are more likely than average to evaluate their texts. Evaluation differences between students due to (L1) language proficiency scores are equally large throughout the L1 writing process. L2 Process Planning is more likely to be carried out by students with higher (L2) language proficiency scores throughout the writing process, but the differences due to language proficiency become smaller towards the end of the writing process.

Figure 2. The effect of language proficiency on L1 Evaluating Own Text, L2 Process Planning, and L2 Content Planning



DISCUSSION

Language proficiency has, in previous research, often been forwarded as a constituent part of writing proficiency. Especially in explanations of L2 writing, which is generally of lower quality, language proficiency has been a prominent factor. However, the effect of language proficiency on both L1 and L2 writing processes has not been uncovered by previous research, because either no writing process data were available (cf. Schoonen et al., 2003), or no individual language proficiency information was available (cf. Chenoweth & Hayes, 2001; Hirose, 2003; Lindgren et al., 2008; Roca de Larios et al., 2001; Van Weijen, 2009), or the research involved L1 writing only (cf. Van der Hoeven, 1997). In the study reported here, both individual language proficiency scores and process data were obtained. The quality of the writing process was operationalized temporally: the moment at which cognitive activities are applied reflects the quality with which the activities are executed. The validity of this temporal approach has been confirmed in a large amount of studies (Breetvelt et al., 1994; Leijten & Van Waes, 2006; Olive et al., 2008; Roca de Larios et al., 2001; Rijlaarsdam & Van den Bergh, 1996; Rijlaarsdam et al., 2011; Van den Bergh & Rijlaarsdam, 1996; Van den Bergh et al., 2009; Van der Hoeven, 1997).

An effect of language proficiency was found for Evaluating Own Text in L1 and for Process Planning in L2. During L1 writing, students with higher (L1) language proficiency scores are more likely than average to evaluate their texts. Evaluation differences between students due to (L1) language proficiency are equally large throughout the L1 writing process. An interaction effect of language proficiency and 'time during the writing process' on L2 Process Planning was established. L2 Process Planning is more likely to be carried out by students with higher (L2) language proficiency scores throughout the writing process, but the differences due to language proficiency become smaller towards the end of the writing process. So, differences between students due to (L2) language proficiency in the amount of (L2) Process Planning applied are most visible at the start of task execution: students with higher language proficiency perform more Process Planning activities than average at the start of task execution, while students with lower language proficiency perform less Process Planning activities than average at the start of the writing process.

In short, language proficiency had an effect on the (temporal) occurrence of a moderate number of cognitive activities: Evaluating Own Text in L1, and Process Planning in L2. Interestingly, the only established effects concern conceptual (Evaluating Own Text, although this activity also has a linguistic component, e.g. the evaluation of spelling and grammar) or regulatory (Process

Planning) activities. Notably, the occurrence of Formulating activities – the most language-specific activity included in the present study – is (unlike in Van der Hoeven's (1997) study) not affected by language proficiency.

There are, however, a number of methodological issues which make it prudent to interpret the results of the reported study with some caution. First, language proficiency was measured by means of timed vocabulary tests. Therefore, the tests may to some degree also have measured retrieval speed, whereas in Schoonen et al. (2003)'s study, retrieval speed made no unique contribution to writing proficiency. In addition, to perform well on this vocabulary test, students need a fair amount of reading competency (i.e. in order to fill in the correct word in the sentence "… *he has a busy life, he always makes time for me*", students need to infer the contrastive relation between 'having a busy life' and 'making time'). In short, the language proficiency tests might have been somewhat diffuse: they might have measured different skills at once, which possibly cancel each other out (i.e. a student has a large vocabulary, but is a poor reader, which might result in an average test score – although vocabulary knowledge and reading proficiency have been shown to be positively related, see Qian, 2002). On the other hand, language proficiency is a diffuse construct, and all of the skills needed to complete the tests (vocabulary size, retrieval speed, reading skill) belong to the domain of language proficiency.

Second, the items in the language proficiency tests were mostly quite generic (common verbs, prepositions, adverbs, et cetera) and not specific tot the topics which the students had to write about. However, it might in future research be interesting to construct vocabulary tests with items which are related to the topic of the writing assignments. Such tests might after all be better predictors of whether students will experience problems with language retrieval during writing.

Third, that high and low language proficient students do not differ in terms of the moment at which they apply most cognitive activities (i.e. other than L1 Evaluating Own Text and L2 Process Planning), does not rule out that there are any other quality differences due to language proficiency in the execution of each of the cognitive activities. For instance, all students evaluate their texts more at the end of task execution than at the start. Nevertheless, the evaluations carried out (at the end of task execution) may differ in nature: For example: are evaluations of the produced text constricted to local concerns, such as the spelling of words and the grammar of sentences, or do the evaluations pertain to the rhetorical structure of the entire text? Or: if evaluations are carried out, do they result in actual improvements of the text? In short, while the temporal distribution of cognitive activities has been shown to be a valid reflection of the quality of the writing process (indeed, the strength of the correlations between predicted and observed values in figure 1 indicates that the temporal model captures a substantial part of writing process quality), the quality of task execution may additionally be reflected in terms of the profundity or correctness with which cognitive activities are applied.

Finally, researchers have suggested the existence of a threshold level (Sasaki & Hirose, 1996; Schoonen et al., 2003): a level of language proficiency which should be attained before students are able to apply conceptual and regulatory activities. This threshold hypothesis has been forwarded in L2 writing research in particular, but it may exist in L1 writing, too (cf. McCutchen, 1996). In the threshold hypothesis, then, language proficiency is a conditional skill for performing writing processes with sufficient quality. It is possible that the participants in the present study had already surpassed this threshold level, so that they were all of them able to apply most of the cognitive activities in a more or less sufficient manner. (Indeed, the between-writer-variance was quite small for most of the cognitive activities, cf. Appendix I. This suggests that students in this population perform their writing processes relatively uniformly.) That (L2) Process Planning and (L1) Evaluating Own Text were the only activities on which effects of language proficiency were found, seems fitting with this explanation. After all, these activities are pre-eminently the kind of conceptual and regulatory activities for which the threshold needs to have been surpassed. This implies that they are difficult to carry out. It is not surprising then, if students are found sufficiently language proficient to execute reading, formulating and revising activities with sufficient quality, but not yet to properly plan and evaluate. In future research, it might be fruitful to investigate the influence of language proficiency in students with lower linguistic ability.

CHAPTER 5

RELATING SELF REPORTS OF WRITING BEHAVIOR AND ONLINE TASK EXECUTION USING A TEMPORAL MODEL

Abstract[8]

Current theory about writing states that the quality of (meta)cognitive processing (i.e. planning, text production, revising, et cetera) is, at least partly, determined by the temporal distribution of (meta)cognitive activities across task execution. Put simply, the quality of task execution is determined more by *when* activities are applied than by *how often* they are applied. *Planning* and *revising* are two extreme writing styles, in which (meta)cognitive activities are temporally differently distributed across the writing process. Planners are writers who generate plans before text production. Revisers use text production as a means to arrive at a content plan. The present study investigates the question whether the online (meta)cognitive processing of secondary school students during writing tasks, as measured by think aloud techniques and keystroke logging, can be predicted by their responses to an offline questionnaire which measures to what degree students considered themselves to be planners and revisers. It was expected that different reported writing styles would entail different temporal distributions of six (meta)cognitive activities: reading the assignment, planning, text production, reading own text, evaluating own text and revising. This hypothesis was partly confirmed. The results show that the online temporal distributions of reading the assignment and planning are different for different degrees of reported writing styles. On the basis of these results, the validity of both the questionnaire and the concept of planner and reviser styles are discussed.

[8] This chapter is a slightly adapted version of: Tillema, M., Van den Bergh, H., Rijlaarsdam, G. & Sanders, T. (2011). Relating self reports of writing behaviour and online task execution using a temporal model. *Metacognition and Learning 6*(3), 229-253.

Writing a coherent and readable text involves handling many (meta)cognitive activities, such as generating, planning, translating ideas into language and making revisions when needed, while keeping rhetorical goals in mind and staying aware of the intended audience (Flower & Hayes, 1980; Hayes & Flower, 1980; Hayes, 1996). The writing process is not a linear chain of actions in which planning, generating, text production and revising, for example, are carried out in consecutive phases. Rather, it is a recursive process, in which cognitive activities may be re-applied during any phase of the writing process (Hayes, 1996). Sometimes a fully developed content plan is subsequently translated into text; in other cases, development of the text plan correlates with the development of the written-down-text. In the latter cases, writing is an act of discovering what to say (Galbraith, 1996; Hayes, 1996). In short, there are numerous possible configurations of (meta)cognitive activities (Rijlaarsdam & Van den Bergh, 1996; Van den Bergh & Rijlaarsdam, 2001; Van Weijen et al., 2008).

Torrance et al. (1994) present two extreme writing styles, with different configurations of writing activities. They identified *planners*, "who planned extensively and then made few revisions" on the one hand, and *revisers*, "who developed content and structure through extensive revision" on the other hand. In addition, they identified so-called *mixed strategy writers*, who applied both planning and revising activities extensively. Similar strategies are found in Torrance et al. (1999; 2000), although they are labeled differently, together with a number of additional strategy types. Biggs et al. (1999) present a typification of writing strategies similar to the planner/reviser distinction. On the one hand, *engineers* plan extensively before commencing with text production. On the other hand, *sculptors* start text production in a relatively early stage of the writing process, without much planning preceding it. The content plan develops as the text develops. The produced text is subsequently revised until it fits what the writer wants to say. Kieft et al. (2006; 2008), finally, also use these two writing styles in their research. They quote Galbraith and Torrance (2004) to describe the *planning strategy* as a strategy "in which writers concentrate on working out what they want to say before setting pen to paper, and only start to produce full text once they have worked out what they want to say" and the *revising strategy* as a strategy "in which writers work out what they want to say in the course of writing and content evolves over a series of drafts" (Kieft et al., 2008, p. 380).

Planners and revisers, then, by definition have different configurations of planning activities, text production activities and revision activities. Other cognitive activities might be expected to have different distributions, too. Reading the assignment, for example, probably occurs more early on in the writing process for extreme planners, but in later stages of task execution for typical revisers. After all, typical planners will think about whether the text matches the assignment during planning stages, while typical revisers will think about this once a text has been produced. The moments of occurrence of all of these activities is, of course, interrelated. If planning happens early on in the writing process, for example, revising cannot.

In short, planners and revisers apply the various (meta)cognitive activities at different moments during the writing process. This leads us to observations made by Rijlaarsdam and Van den Bergh (1996) and Van den Bergh and Rijlaarsdam (1996). They demonstrated that the occurrence of (meta)cognitive activities varies across task execution. Structuring activities (a subcomponent of planning), for example, are on average more likely to occur a short while after the start and also towards the end of task execution, but less likely to occur during middle stages of the writing process. They also showed that the distributions of cognitive activities differ between individual writers. Some writers, for example, follow the average distribution of structuring activities, while others tend towards a different distribution. One, for example, in which structuring activities are hardly used at the start of the task, a little more during middle stages, and mostly during the final phases of the writing process. Finally, Rijlaarsdam and Van den Bergh (1996) and Van den Bergh and Rijlaarsdam (1996) demonstrated that the relation between cognitive activities and text quality varies across task execution. Structuring activities, for instance, were shown to be more effective when they occurred during early stages of task execution (i.e. the correlation between structuring and text quality is at its highest at the start of the writing process) and less effective when they occurred towards the end of task execution. They therefore advocated a temporal analysis of activities over the writing process: analyses of cognitive processing during writing should take the moment(s) at which cognitive activities occur into account. The validity of this temporal approach was confirmed by Breetvelt et al. (1994), Van den Bergh and Rijlaarsdam (2001), Van den Bergh et al. (2009) and Van Weijen et al. (2008), who also demonstrated that differences between writers in terms of distributions of cognitive activities explain differences in the quality of the texts produced. The temporal approach of writing processes has become a dominant view in writing research (Leijten & Van Waes, 2006; Olive et al. 2008).

The configuration and temporal distribution of (meta)cognitive activities is an online characteristic: we can establish it by measuring what happens during the process of task execution. A common method for measuring the process of task execution (during writing, but also during other tasks, for example reading tasks or mathematic problems) is the use of think aloud techniques (cf. Cromley & Azevedo, 2007, Rijlaarsdam & Van den Bergh, 1996; Roca de Larios et al., 2008;

Van den Bergh & Rijlaarsdam, 1996; Van Weijen et al., 2008; Veenman & Spaans, 2005). Another method for concurrent measurements of writing processes is keystroke logging (e.g. Leijten & Van Waes, 2006; Strömqvist, Holmqvist, Johansson, Karlsson & Wengelin, 2006). Online measurements have been shown to have predictive value for the quality of the output of task execution (Cromley & Azevedo, 2007; Torrance et al., 1999; Van der Stel & Veenman, 2008; Van Weijen et al., 2008; Veenman, Prins & Verheij, 2003). This output may be text quality (for writing), but also test scores (in other domains, such as reading and mathematics).

However, there are also numerous studies where writing behavior is measured independently from the writing process. Questionnaires about different aspects and/or configurations of writing processes (Kieft et al., 2006, 2008; Lavelle, Smith & O' Ryan, 2002; Torrance et al., 1994, 1999, 2000) are an example of such offline measures. The use of offline measurements has been criticised as inaccurate reflections of the underlying process. Russo, Johnson and Stephens (1989), for example, found the contents of retrospective protocols to be incomplete and partly fabricated. Their opinion is shared by Veenman et al. (2003) and Cromley and Azevedo (2006). In both studies, offline reports were related to online data, the latter in the form of total or relative frequencies of strategy-related verbalizations in concurrent data (Cromley & Azevedo, 2006; Veenman et al., 2003) or by proportions of indicated strategy use in a concurrent multiple-choice tool (Cromley & Azevedo, 2006). They found that relations between offline reports and online task execution were weak or absent.

However, analyzing online metacognition by establishing frequencies of metacognitive verbalizations runs counter to the idea that the quality of online task execution is determined by the *temporal distribution* of (meta)cognition across the writing process. Possibly, this could form an explanation for the absence of (substantial) relations between offline and online data in these studies.

Torrance et al. (1999) indeed showed a correspondence between questionnaire outcomes and online data, the latter being analyzed in terms of distributions. Participants in their study completed a questionnaire about their writing behavior. On the basis of this questionnaire, participants were categorised into one out of three possible strategy groups. The questionnaire outcomes in this study predicted online writing behavior. This online behavior was analyzed in terms of distributions of (meta)cognitive activities, such as planning, translating ideas into language and revising. Torrance et al.'s (1999) study, then, suggests that offline reports of writing behavior[9] can, at least to some extent, be used as predictors of a general tendency towards a particular online distribution or configuration of cognitive activities.

A Writing Style Questionnaire developed by Kieft et al. (2006; 2008) measures reported degrees of planner- or reviser-type writing behavior within individuals. Contrary to Torrance et al. (1999), Kieft et al. (2006; 2008) do not categorize writers, i.e. writers are not either planners or revisers. Rather, the two dimensions (i.e. the planner and reviser dimension) are seen as scales, on both of which the degree to which it applies to an individual writer is expressed. Kieft et al. (2008) provide some evidence which seems to suggest a degree of validity for this questionnaire. They tested students' writing style by means of the Writing Style Questionnaire. Subsequently, all students participated in a lesson series on writing. One group of students (consisting of both students for whom planning was the dominant writing style and students for whom revising was the dominant writing style) received instruction that matched a planning style and another group of students (again consisting of both students for whom planning was the dominant writing style and students for whom revising was the dominant writing style) received instruction that matched a revising style. They found that study outcomes (i.e. the quality of the texts which the students wrote) increase if writing lessons match the most dominant writing style in students' responses to the questionnaire. Kieft et al.'s (2008) result is, however, indirect evidence for the assumption that the Writing Style Questionnaire is a predictor of online writing behavior. There has, to date, been no research to test whether higher or lower degrees of reported planner- or reviser-type behavior do indeed predict different online configurations and temporal distributions of (meta)cognitive activities.

This is investigated in the present study. It may be assumed, for example, that 'high planners' (according to the Writing Style Questionnaire) perform more planning activities at the start of task execution than 'low planners'. After all, we know that planning activities are most effective during initial stages of the writing process (Van den Bergh & Rijlaarsdam, 2001). Similarly, it may be assumed that 'high planners' will apply less planning activities than 'low planners' at stages in the writing process during which planning activities are less effective: towards the end of task execution. 'High revisers', on the other hand, will generally apply more planning activities at the end of task execution than 'low revisers'. After all, typical

[9] Offline reports of writing behavior are essentially reports of what learners know about their writing activities/strategies. They are, in other words, measures of what Flavell (1979) calls *metacognitive knowledge about strategy,* or what Zohar and David (2009) call *Meta-strategic Knowledge*: "an awareness of the type of thinking strategies being used in specific instances".

revisers use text production to arrive at a plan of what to say. As a consequence, ‘high revisers’ will also apply more revision activities at the end of task execution than ‘low’ revisers. In the same vein, we predict that different scores on the planning and revising dimension in the Writing Style Questionnaire (Kieft et al., 2006, 2008) are related to different distributions (across task execution) of the other (meta)cognitive activities which occur during writing.

METHOD

Participants

The participants were fourteen- and fifteen-year-old students (N = 20; 10 female and 10 male). They were from three different third-year-forms at the same school for pre-university secondary education. They were recruited by means of a call for volunteers, which was distributed by their Dutch language teacher. All participants were native speakers of Dutch. They received a small financial compensation for their participation. Parental consent was obtained.

Tools and procedures

The students completed four writing tasks. In addition, they completed an offline questionnaire to measure reported writing behavior. They performed all tasks individually in a university room, in the presence of a test leader.

Writing tasks

All students wrote four argumentative essays in Dutch, their mother language, on topics such as 'camera surveillance in inner city areas' or 'legalisation of soft drugs'. They completed all their essays during one session, with a short break of about fifteen minutes between assignments. The sequence of topics was systematically balanced across participants.

The assignments consisted of a brief statement of topic, audience (peers), medium (the school paper) and purpose (to convince the readers of your point of view), followed by a series of quotes (factual information as well as opinions) that were related to the topic, of which two had to be used in the essay. All assignments were tested with third year students of pre-university secondary education during a pilot study in 2005. They were also successfully used by Van Weijen et al. (2008) and Van Weijen, Van den Bergh, Rijlaarsdam and Sanders (2009). The essays had to be about half a page in length (which is about 250 to 300 words). An example of an assignment can be found in Appendix A.

The available time for each essay was thirty minutes. The mean writing time was 20.13 minutes (*SD* = 5.89, *Min.* = 7.80, *Max.* = 32.15). The time spent on each task was related to the order in which tasks were completed ($\chi^2 = 7.23$, $df = 1$, $p < .05$). The mean writing time for the first essay in the session was 21.74 minutes, while the mean writing time for the last essay was 18.93 minutes. That students spent less time on the last task in the session than on the first task can probably, for a large part, be explained by the fact that students generally needed less time for reading the assignment during later tasks: a fairly large portion of the instruction text (e.g. description of audience, medium and purpose) was identical in all tasks.

The students wrote the essays on a computer using Microsoft Word. They had to think aloud during the process of task execution. All writing sessions were video-taped. The writing sessions were also recorded by means of keystroke logging (Inputlog: Leijten & Van Waes, 2006), in order to obtain more detailed information on text production and revision activities.

Writing style questionnaire

The students also completed Kieft et al.'s (2006; 2008) Writing Style Questionnaire. This questionnaire measures reported degrees of planning and revising style. It is specific to the domain of argumentative writing, in that participants are asked how they would handle writing an argumentative essay about the tobacco industry. This 'tobacco' task, which was not actually carried out by the participants, is very similar to the four writing tasks performed by the students in the present study in terms of text type and intended medium.

The questionnaire consisted of thirty-six statements about writing strategy. Thirteen of these items reported planning-type behavior and twelve of these items reported revising-type behavior. The remaining eleven items are fillers. Students had to indicate in how far each statement pertained to them, by checking a box on a five-point scale. On the basis of their questionnaire responses, participants received scores for both the planning dimension and the revising dimension. They could therefore score equally high or low on both dimensions, or one of the two dimensions could be dominant. Figure 1 features all questionnaire items, which are sorted according to the dimension they pertain to. In the actual questionnaire, the items were presented in random order. For the present study, the questionnaire was in Dutch, the students' mother language.

Analyses

All think aloud data were transcribed and segmented, which were completed by the Inputlog recordings. A new segment in the protocols reflected a switch to a different (meta)cognitive activity within a participant’s writing session (Van den Bergh & Rijlaarsdam, 2001).

*Figure 1. Items in the Writing Style Questionnaire (Kieft et al., 2006; 2008), sorted according to which dimension they measure. *: item is negatively formulated*

Planning

Before I start writing, I want to have it clear which information to put in the text.

Therefore, planning is important to me.

If I have to write a text, I spend a lot of time on thinking about my approach.

I always make a text schema before I start writing.

If I have to write something, I jot down some notes, which I work out later.

Before I start writing a text, I write something on a scribbling pad, to find out my opinion about the topic.

* Planning is of no use to me.

* When I start writing, I don't yet have a clear idea of what will be in the text.

Before I start writing, I have a clear picture of what I want to achieve with the readers.

I need to have my thoughts clear before I am able to start writing.

Before I write a sentence down, I already have it in my head.

* When I am writing, I sometimes write down pieces of text of which I know that they are not completely right yet. Still, I prefer to go on writing at that point.

* When I read over my texts, I usually find a lot to improve.

* When I read over my texts, they are sometimes very chaotic.

Revising

* I always start writing straight away: I don't need to know exactly what I will write or how the text will be built-up. That will become clear as I write.

When my text is ready, I read it through thoroughly and make improvements: a lot can still be changed at that point.

During writing I regularly check if my text does not contain any sentences which are incorrect or too long.

While writing my text, I continually ask myself if readers will be able to follow it.

For me, writing is a way to get my thoughts clear.

* I usually hand in my text without checking if its organization is in order.

*Figure 1 (continued). Items in the Writing Style Questionnaire (Kieft et al., 2006; 2008), sorted according to which dimension they measure. *: item is negatively formulated*

If I read over my texts, and rewrite my texts, it occurs regularly that I drastically change their organization.

Before I hand in a text, I always check if its build-up is logical.

* I never pay much attention to whether I have forgotten to put any sentences or ideas in a text.

When I rewrite a text, the content usually changes drastically, too.

When I finish a text, I usually need to read through it carefully, to check if there is no superfluous information in it.

I never pay much attention to whether I am satisfied with my texts.

Fillers

I write and rewrite my text sentence per sentence. Only if I am completely satisfied with a sentence, do I proceed with writing.

When I am writing, I find it hard to organize my thoughts.

Only if my text is complete, do I read what I have written.

If finally I have an approximate idea of what to say in my text, the words will flow out of my pen.

When I write, I stop writing after every few sentences to read what I have just written.

I try to write a correct version of my text in one go, so that I hardly have to make any alterations when it's finished.

When I write a text, I find it hard to come up with ideas.

When I am writing, I often find that all kinds of new ideas pop into my head.

For writing tasks, I do not find it very hard to think of arguments to support my point of view.

The texts which I write are usually not very original.

I make sure that every sentence is perfect, before I start with the next sentence.

When my text is finished, the only thing I do is check for language or spelling mistakes.

All segments were coded according to a coding scheme (adapted from Breetvelt et al., 1994). One think aloud protocol (number of segments = 518) was coded by two researchers. The intercoder agreement (Kappa) was 0.85. The coding scheme (see table 1) consists of fourteen categories. Five of these categories reflect planning activities, namely monitoring, goal setting, generating content, structuring (which involves the selection and evaluation of propositions which have been generated but not (yet) translated into text) and metacomments. One category involved subcodings, namely ’Revising’. Revisions were subcoded as ‘automated corrections’ when they involved corrections of typographic errors. They were typically errors which are made as a result of the use of a keyboard, which seem to be corrected almost automatically. An example would be: a writer types ‘almots’, and immediately corrects this error by giving to backspaces and typing ‘st’, so that it now reads ‘almost’. All revisions which were not ‘automated corrections’ were subcoded as ‘conceptual revisions’: they involve alterations which are made by the writer in a non-automated way. That is, they

Table 1. Coding categories in the coding scheme

	Coding category	**Description**	**Example**
READING THE ASSIGNMENT	Reading the instruction text and documentation	Reading (part of) the task instructions	"Write an essay in which you..."
PLANNING	Monitoring	Verbalizations of writing process management. Mostly self-instructions	"I'm going to read what I've written so far."
	Goal setting	The formulation of goals which the text has to satisfy	"The text should be convincing."
	Generating	Generating ideas for content or form	"Something about the disadvantages of camera surveillance..."
	Structuring	Evaluating and arranging ideas	"Something about adults? No, that's not relevant."
	Metacomments	Evaluations of a student's own writing process	"I should have made an outline of the text before I started."

Table 1 (continued). Coding categories in the coding scheme

	Coding category	Description	Example
TEXT PRODUCTION	Text production	Production of new text	"Camera surveillance invades people's privacy." (Verbalizations usually occur parallel to the activity of typing.)
READING OWN TEXT	Reading produced text	Reading (part of) the produced text, at any given moment during the writing process	" Surveillance cameras do not increase public security."
EVALUATING OWN TEXT	Evaluating produced text	Evaluation of produced text	"The largest part is about backdraws."
REVISION	Revising	Making changes to the text produced so far at word, sentence or text level.	Moving a set of sentences from the body of the text to the introduction.
	Automated corrections	Corrections of typographic errors due to keyboard use. *(Not included in the analyses.)*	Writer types 'almots', deletes 'ts' by means of 'backspace', en then types 'st'. (Mostly no (explicit) verbalization of the correction.)
OTHER	Pauses	Silence or interjection	"eeeerrr"
	Interaction with test leader	Interaction between test taker and test leader	"Could I open a window?"
	Physical activity	Physical activity	Taking a sip of tea.
	Navigation	Moving through the document: arrow buttons or mouse movements	Moving the cursor some lines back.

involve actual alterations at the level of spelling or content. In the remainder of this article, we will mean conceptual revisions if we use the term 'revision' or 'revising'.

As different degrees of reported planner- or reviser-style behavior entail different configurations of the complete writing process, we expect different distributions for all of the activities listed in table 1: reading the assignment, text production, planning, reading own text, evaluating own text, and revision. An exception is formed by the activities categorized as 'OTHER'. For these activities, there is no conceptual link with planner and reviser styles. The same applies for the subcategory of 'automated corrections'. These five activities (pauses, interactions with test leader, physical activity, navigation and automated corrections) were therefore not included in the analysis.

Figure 2 shows an example of (a part of) a protocol and illustrates how the think aloud data en the Inputlog data were integrated. All information in the column labelled 'Typing' are derived from Inputlog. This column contains 'text production' and 'revising' activities. The protocols consisted of seven columns. The column labeled 'Reading' was used for indicating if any reading activities (reading the assignment - RA – or reading own text – ROT) were taking place. The 'Verbalizations' column contained everything which was said out loud by the student, except interjections, such as "uhm". The 'Typing' column contained all text production as registered by Inputlog. There were three categories in this column, namely the production of new text, revisions (indicated by Inputlog as [BS] for backspace or [DEL] for if the delete button was pressed), and navigation (by means of mouse movements or arrow buttons). The 'Pausing' column contained all silences and interjections. The column labeled 'Other' mostly contained descriptions of physical activities, for example 'takes a sip of his drink'. One row is one protocol segment. As such, this transcription method allows for parallel

Figure 2. Part of a completed protocol

Segment number	**Reading**	**Verbalizations**	**Typing**	**Pausing**	**Other**	**Code**
17		even een titel erboven (*I'll insert a title)*				Monitor
18		having	Having childere			Text production
19		chil	[BS 1] [BS 1] [BS 1]			revision (automated correction)
20		children yes or no	ren, yes or no?			text production
21	ROT	having children yes or no				Reading own text

actions. Text production and verbalizations, for example, often occur simultaneously. The 'Typing' column would contain the text production as registered by Inputlog. The codings in the last column were in reality numbers. Code 01, would stand for 'reading the assignment', for example, and code 02 would stand for monitoring. English translations of Dutch verbalizations are given in italics.

Due to technical deficits, there were less than four writing sessions available for analysis for three participants. For one participant, two writing sessions were included in the analysis. For two other participants, three writing sessions were included. For the remaining seventeen participants, all four writing sessions were available.

Modelling (meta)cognitive activities across task execution

The first step in the analysis is to model the (online) occurrence of (meta)cognitive activities temporally, that is, as a function of the moment in the writing process. We constructed this time variable by splitting each protocol into five equally long episodes in terms of numbers of segments (cf. Roca de Larios et al., 2008; Van den Bergh & Rijlaarsdam, 1996). A protocol of 330 segments, for example, would be analyzed as five episodes consisting of 66 segments each. By using episodes we achieved standardisation: it allowed us to compare different writing processes between (and within) individuals in terms of start, middle and end of task execution. After all, episode 3, for example, reflects the middle part of the writing process for each protocol, no matter if it contains segments 133 to 199 in a protocol of 330 segments, or segments 101-150 in a protocol of 250 segments. The (meta)cognitive activities which are the dependent variables in our analysis (i.e. reading the assignment, planning, text production, reading own text, evaluating own text, and revision) were all expressed as proportions of the total number of segments for each episode. For instance, if an episode consisted of 80 segments (which would mean that the entire writing process consisted of 400 segments), and 10 of these segments were coded as 'reading the assignment', then the proportion for reading the assignment in that episode would be 0.125.

A multilevel regression model has been applied to model the occurrence of the (meta)cognitive activites at each episode, as episodes are nested within writers (Van den Bergh et al., 2009). The analysis was conducted with MLwiN software for multilevel models. In effect, a longitudinal model is in operation, as it concerns changes in occurrence during the writing process: proportions of the applied (meta)cognitive activity may be different during each new episode. Therefore, the occurrence of each of the six online activities (A) had to be described as a function of episode, i.e. A=*f*(episode). Note, however, that this function *f* does not need to be identical for all individuals *i*: A=f_i(episode).

This function, *f*, can take many forms (Goldstein, 1979; Healy, 1989). For this study, polynomial models were preferred because of their flexibility. Depending on the number of coefficients (and their numerical values), polynomials can take almost any shape. As such, they can be used to model various kinds of growth patterns.

Growth across task execution is not necessarily linear. For instance, text production activities may occur relatively little during first and last episodes of the writing process, but a lot during the middle part of task execution (e.g. during episode 3) Therefore, non-linear terms (e.g. quadratic or cubic terms) can also be included in the model: the occurrence of an activity (at each episode) is described as powers of episode (episode^0, episode^1, episode^2, …). The number of parameters needed to describe the observed activities (in each episode) is considered an empirical matter. That is, a next power of episode is only included in the model if it has a significant contribution in the description of an activity *and* if all lower powers are significant as well (see, Van den Bergh & Rijlaarsdam , 1996). For example, 'episode' to the second power can only be added to the model, if the linear term (episode^1) has reached significance.

To meet the requirement that the function *f* is allowed to differ between individuals, not only are the regression coefficients of the powers of episode estimated, but also the variance of these parameters. That is, the variance of the intercept (writers differ in occurrence at episode = 0), the variance of the linear component (writers differ in linear change over the writing process), et cetera.

These variance components are in fact the variances of residuals which characterize the occurrence of activities of a specific writer. Therefore, the differences between individuals can be explained by individual characteristics like their offline planner and reviser scores. Adding these offline scores, then, is the final step in the construction of a multilevel regression model, in which episode and the individual planner or reviser scores were the explanatory variables. Of course, the effect of these offline scores is not (necessarily) constant across task execution. (For instance: we expect differences in process execution due to higher planner scores to be larger at the start of task execution than during later parts.) Therefore, interaction effects between the offline scores and the time variable (episode) on the dependent variable were also calculated. The complete multilevel regression model, as construed in MLwiN and as used for explaining the occurrence of each of the six (meta)cognitive activities, can be found in Appendix K.

RESULTS

Internal consistency was calculated for the Writing Style Questionnaire. Cronbach's alpha is .72 for the items on the planner dimension and .64 for the items on the reviser dimension. These reliabilities, which are similar to the reliabilities found by Kieft et al. (2006; 2008), justify aggregating the items for each dimension to calculate mean scores per dimension per student. As the two dimensions are only very moderately correlated (r = .39), planning- and revising-type behavior can be identified separately in the Writing Style Questionnaire data (see also Kieft et al., 2006; 2008).

Table 2. Descriptive statistics for the Writing Questionnaire scores, and (meta)cognitive activies in the concurrent protocols (absolute numbers and proportions). Means (M)*, standard deviations* (SD)*, minimum and maximum scores/numbers of verbalizations* (Min.; Max.)

Descriptives				
	M	*SD*	*Min.*	*Max.*
Writing Style Questionnaire				
score *planning* dimension	2.45	.66	1.50	4.13
score *revising* dimension	3.16	.57	2.25	4.38
numbers of segments per participant, per task				
total	346.68	137.88	134	767
reading the assignment	18.96	14.48	5	75
planning	19.21	19.71	0	96
text production	131.16	56.77	44	282
reading own text	11.18	11.65	0	54
evaluating own text	2.39	3.21	0	13
revising	0.20	0.65	0	4
proportions of each (meta)cognitive activity per episode				
reading the assignment	.0569	.05793	.00	.31
planning	.0544	.06143	.00	.40
text production	.3791	.09192	.10	.58
reading own text	.0295	.04004	.00	.25
evaluating own text	.0061	.01636	.00	.15
revising	.0005	.00367	.00	.04

Table 2 features descriptive information about the data. It shows that writing processes are on average 346.68 segments long, but that there is great variation (*SD* = 137.88). In addition, table 2 shows that text production, which on average takes up about 131 segments (which is about 38% of the segments), occurs far more often than the other activities in the analysis, which on average occur in 0,5% to 5% of the segments[10]. However, the data show a relatively larger range for the more infrequent activities than for text production. The standard deviations for 'reading the assignment', 'planning', 'reading own text', 'evaluating own text' and 'revising' are in most cases larger than their means. For these activities then, there seems to be a lot of variation due to episode (i.e. moment during the writing process) and student. This variation is, of course, to be explained by the results of the regression analyses.

The results of the regression analyses show that for the six online activities which were analyzed, the average occurrence indeed varies across task execution, i.e. due to 'episode'. (See Appendix L for parameter estimates.) The proportion with which reading the assignment, planning, text production, reading own text, evaluating own text and revising are applied, is significantly different if episode is the explanatory variable.

Pseudo R^2 was calculated for the six models which were constructed to explain the occurrence of online activities with the 'episode' variables. The outcome is presented in table 3. Only for one of the activities, R^2 is low (reading own text: $R^2 = 0.18$). As this is not a crucial activity in our analysis, this is a relatively minor problem. For the other five activities (reading the assignment, planning, text production, evaluating own text and revising), R^2 proved to be satisfactory.

Table 3. The fit of the constructed regression models (R^2)

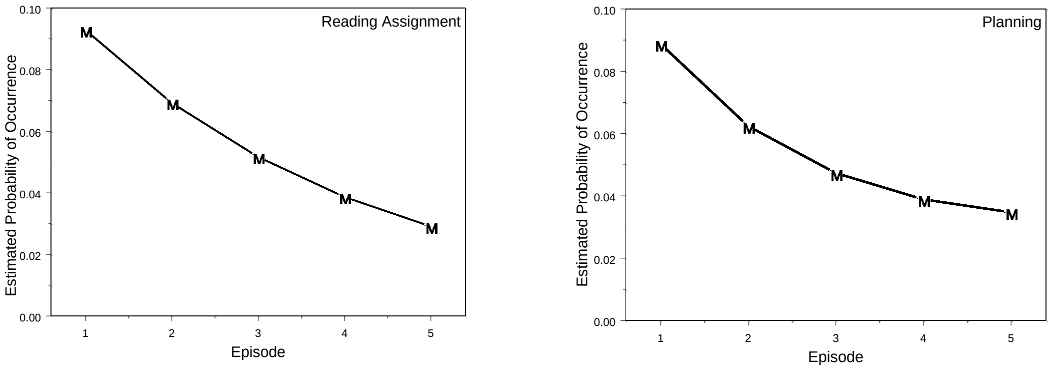
Activity	**R^2**
Reading the assignment	.69
Planning	.59
Text production	.58
Reading own text	.18
Evaluating	.78
Revising	.81

Figure 3 (a, b, and c) shows the average distributions of all six activities. The relation between online activities and episode was analyzed in logits. As logits are hard to interpret, we transformed them into proportions, in order to interpret the

[10] 'Pausing' and 'automated corrections' were relatively frequently occurring activities. This explains why the accumulated proportions for the six (meta)cognitive activities presented in table 2 do not approximate 1. (.0569 + .0544 + .3791 + .0295 + .0061 + .0005 = 0.497)

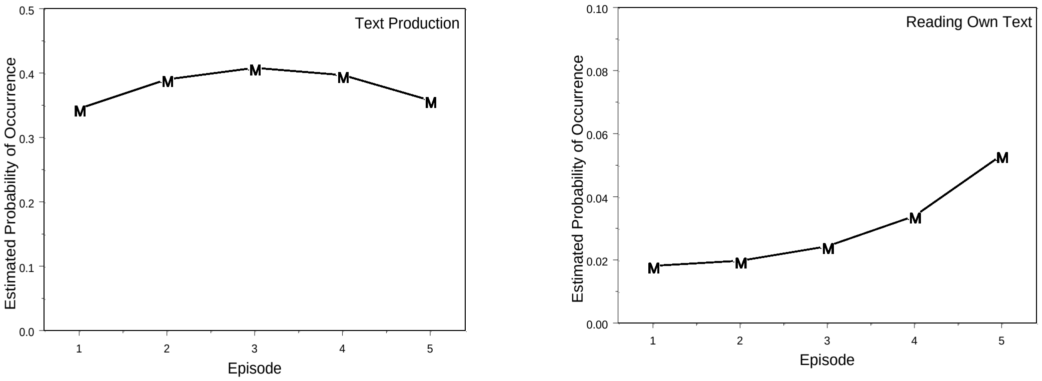
results as probabilities of occurrence. In the figures below, then, we present distributions of activities in proportions.

Figure 3a. Average distributions across task execution for Reading the Assignment *and* Planning



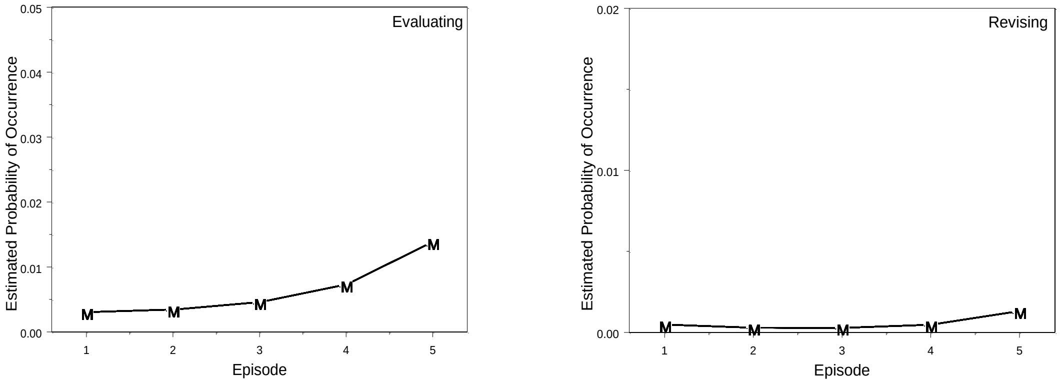
The probability that 'reading the assignment' occurs is highest in episode 1, that is, at the start of task execution. Thereafter, its probability of occurrence declines (with slightly different amounts) with every next episode. To put it more simply: reading the assignment happens most often at the start of the writing process (episode 1), and least often at the end of task execution (episode 5). The same pattern applies for planning activities.

Figure 3b. Average distributions across task execution for Text production and Reading Own Text



Text production activities are distributed differently across task execution. They are already quite likely to occur at the start of task execution, with an estimated probability of almost .35. After the start of task execution, the probability increases, reaching its peak at the middle stage of the writing process. After episode 3, there is a slight decrease towards the end of task execution. Text production activities, in short, occur quite frequently across the entire writing process, but are most frequent during the middle part of task execution. Reading own text occurs least at the start and most at the end of task execution, although its likelihood of occurrence is still very low at the end of the writing process.

Figure 3c. Average distributions across task execution for Evaluating Own Text and Revision



Evaluating own text occurs least at the start and most at the end of task execution, although its occurrence is still unlikely at the end. This is even more the case for revision activities. Although they are very slightly more probable during episode 5, revision activities on average occur infrequently at any moment during writing. The next question now is if these distributions vary according to offline planner and reviser scores. Table 4 gives an overview of whether significant effects were found. The second and third column show if there is a significant main effect of planner or reviser scores on the number of activities applied. Such an effect would mean that differences in offline scores are related to differences in the number of times that an activity occurs during the entire writing process. The fourth and fifth column show if there are significant interaction effects of planner/reviser and episode. The existence of such an effect would mean that the effect of higher or lower offline scores varies across episodes. It would mean, in other words, that distributions are different for different planner or reviser scores. (See Appendix M for parameter estimates.)

Table 4 shows that there is a main effect of planner scores on four of the six online activities, namely reading the assignment, planning, text production and revising. Significant main effects of reviser scores exist for the activities planning and reading own text. Interaction effects could be established in two cases: 1) for online planning activities, the effect of planner scores is different for various episodes, and 2) for reading the assignment, the effect of reviser scores is different for various episodes.

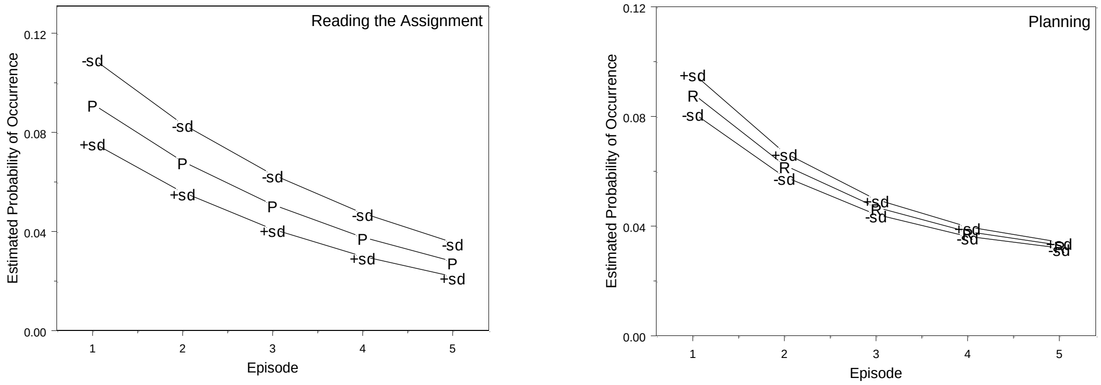
Table 4. Overview of the effects of offline planner and reviser scores and episode on six online (meta)cognitive activities. V: a significant effect was observed

	Planner	**Reviser**	**Planner*episode**	**Reviser*episode**
Reading the assignment	V			V
Planning	V	V	V	
Text production	V			
Reading own text		V		
Evaluating own text				
Revising	V			

The direction of the effects can be inferred from the regressions weights, and are illustrated in figure 4. Again, the logits were transformed into proportions to facilitate the interpretation of the graphs. Figure 4 (a, b, c, and d) shows the variations in the occurrence of (meta)cognitive activities according to variations in offline planner and reviser scores. Each graph contains three lines: one (P or R) reflecting the occurrence of the activity for students with average planner or reviser scores, one (+sd) reflecting the occurrence for students with a planner or reviser score of one standard deviation above the average score, and one (-sd) reflecting the occurrence for students with a planner or reviser score of one standard deviation below the average score. These figures are based on parameter estimates from models in which the effects of planner and reviser scores are estimated simultaneously. The graphs for planner effects therefore assume mean scores on the reviser scale (*z score* reviser = 0), while the graphs for reviser effects assume mean scores on the planner scale.

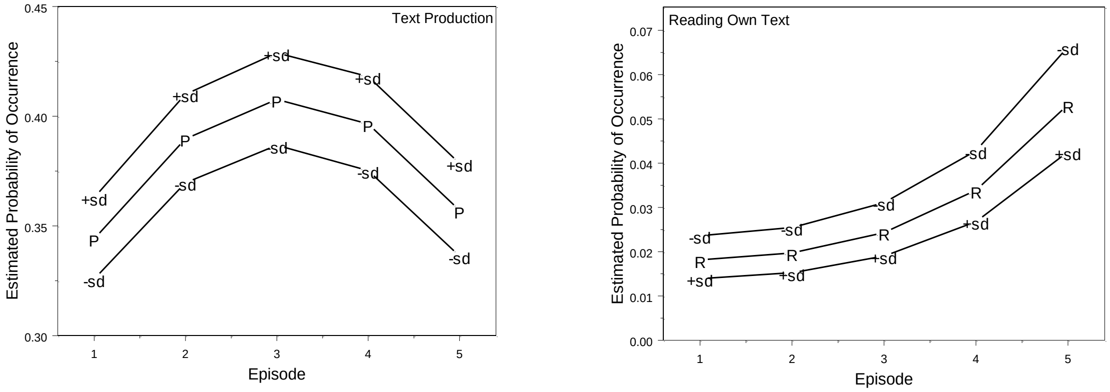
Figures 4a, 4b, and 4c illustrate the established main effects (cf. table 4). Although in these cases the (significant) effects of planner of revisers scores were stable across task execution (i.e. no significant interaction effects of planner or reviser scores and episode could be established), figures 4a, 4b, and 4c illustrate the established effects as observed throughout the writing process.

Figure 4a. Main effects of planner scores on reading the assignment (left) and of reviser scores on planning (right)



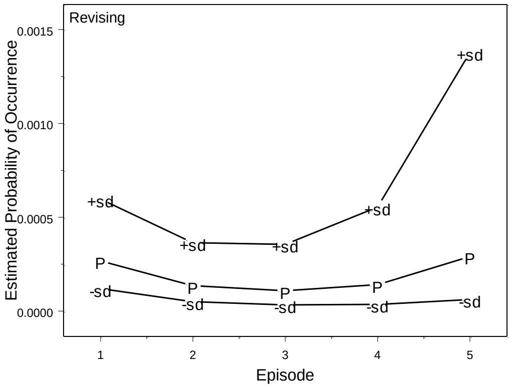
Students with a higher planner (+sd) score generally read the assignment on fewer occasions than average. Students with a lower planner (-sd) score read the assignment more frequently than average. Although the difference between high and low planners seems to become smaller as the writing process progresses, there was no significant interaction effect of planner score and episode. In other words, no evidence could be found that the effect of a higher or lower planner score is different for different episodes. We have to assume that it is stable across task execution. Surprisingly, students with higher (+sd) reviser scores are more likely than average to engage in planning activities, and students with lower reviser scores are less likely to plan. This effect holds throughout the writing process.

Figure 4b. Main effects of planner scores on text production (left) and of reviser scores on reading own text (right)



For text production, there was a significant main effect of planner scores: the higher the offline planner score, the more text production activities occur. For reading own text, there was a significant main effect of reviser scores: the higher the offline reviser score, the less 'reading own text' occurs.

Figure 4c. Main effect of planner scores on revising.



For revision activities, the difference between high and low planners seems larger at the start and the end of task execution. This interaction effect was, however, not significant. There was a significant main effect of planner scores on revising, in an unexpected direction: the higher the offline planner scores, the more revising activities occur.

Figure 4d. Effects of planner scores on the distribution of planning (left) and of reviser scores on reading the assignment

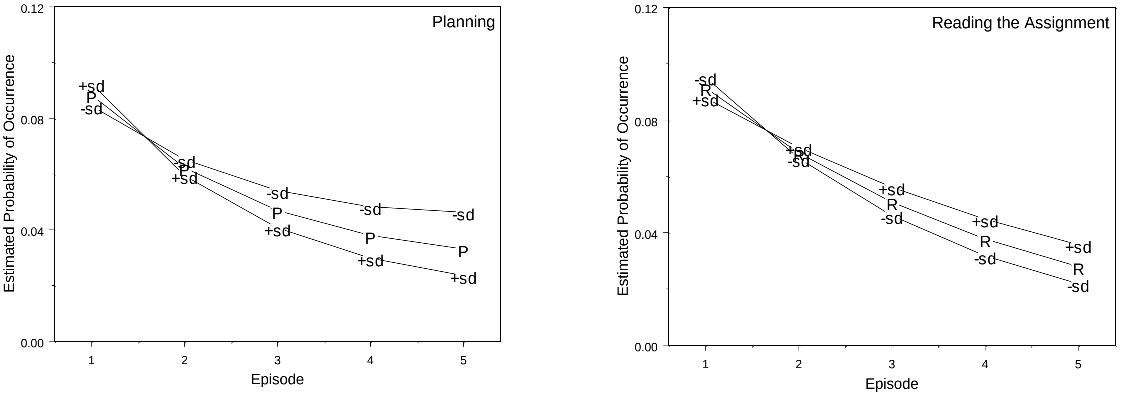


Figure 4d illustrates the established interaction effects. For planning activities, the effect of higher or lower planner scores varies across task execution, i.e. there is a significant interaction between the variable ‘planner score’ and the variable ‘episode’. Students with higher planner scores (+sd) are more likely to apply planning activities at the start of task execution than students with lower planner scores. This changes fairly soon after the start of task execution, so that towards the end of task execution, students with higher planner scores are less likely to apply planning activities than students with lower planner scores. The change in the occurrence of planning activities is therefore stronger for students with higher planner scores. The effect of reviser scores on the occurrence of reading the assignment also varies over time. At the start of task execution, students with lower reviser scores (-sd) are (slightly) more likely to read the assignment than students with higher reviser (+sd) scores. From episode 2 onwards, however, this effect is reversed. From that moment on, students with lower reviser scores are less likely to read the assignment than students with higher reviser scores. The difference (in terms of the probability that ‘reading the assignment’ occurs) between students with higher and lower reviser scores grows larger towards the end of task execution.

DISCUSSION

We predicted that differences in reported planner and reviser styles as measured by Kieft et al.'s (2006; 2008) Writing Style Questionnaire were related to different distributions of various (meta)cognitive activities over the course of the writing process. Results indicate that the occurrence of six (meta)cognitive activities - reading the assignment, planning, text production, reading own text, evaluating own text and revising - varies across task execution. Activities are more likely to occur during some episodes than during others. Different distributions due to reported writing style were found for two out of the six activities which were analyzed, namely for reading the assignment and planning. For three other activities, namely text production, reading the assignment and revising, the effect of different degrees of reported planner or reviser styles did not vary across task execution, but a main effect of planner and reviser scores was established nonetheless: the higher the offline score, the more (or less) frequent do these activities occur during task execution.

The variation in distributions found for planning activities (cf. Figure 4d, left) fits the available theory about the planner style. Students who report a higher degree of planner-type behavior apply more planning activities at the start of task execution, but less towards the end of the writing process. This is in line with the idea that planners do most of their planning before they write anything down. The variation in distributions found for reading the assignment (cf. Figure 4d, right) also fits the available theory about the reviser style. Students who report a higher degree of reviser-type behavior read the assignment more often towards the end of task execution. This makes sense, because revisers think about a content plan during and after text production, that is, during later stages of task execution. It follows that typical revisers will also mostly think about the match between the produced text and the assignment during these later stages.

Various explanations come to mind for the fact that different distributions due to differences in reported writing styles could not be established for text production, reading own text and revising (i.e. there were main effects, but no interaction effects), and no effects were found at all for evaluating own text. One explanation is that, except for text production, these are low-frequent activities. The second explanation has to do with the nature of the activities in this specific age group. This seems to pertain particularly to text production, which is a frequent activity. Although its probability of occurrence is different in different episodes, this variation between different moments in the writing process is quite a bit smaller than it is in more developed writers (cf. Van Weijen, 2009, who used the same assignments in a group of first year university students and found that text production activities were not likely to occur at all at one stage of task execution, but very likely to occur during other stages; temporal variation was, in short, much larger). There is, in other words, less variation to explain in the first place.

There were two seemingly surprising main effects, namely that students with higher reviser scores planned more than average, and that students with higher planner scores revised more than average. Although this seems illogical at first sight, we should keep in mind that the planner and reviser dimensions are not each other's reverse: high scoring planners are not automatically low scoring revisers, and high scoring revisers are not automatically low scoring planners (cf. the low correlation between the Writing Style Questionnaire's planner and reviser dimension, p. 95). In addition, the planner/reviser theory does not stipulate that revisers do not plan. Indeed, they are expected to plan, but later in the writing process than typical planners.

It is striking that there were fewer effects – namely three – due to reported reviser behavior than due to reported planner behavior – namely five. Although it is possible that this is a chance finding, taken together with the fact that the reviser dimension in the Writing Style Questionnaire had lower reliability than the planner dimension (.64 versus .72), this seems to raise some doubts as to the usability of the reviser dimension for less proficient writers, such as the participants in the present study. This idea is supported by the observation that revising, and also reading and evaluating own text, which are associated activities, are extremely low-frequent activities in this age group. In addition, it might be the case that the definition of revisers in the Writing Style Definition is not that clear-cut. It seems that two definitions are simultaneously in operation: one which focuses on the tendency to rely on revision, and one which focuses on how revisers use text production as a means to arrive at a content plan. Actually, the tendency to revise might be a side-effect of revisers' use of text production to get an idea of what they want to say. After all, their initial text production serves planning purposes and the resulting text is therefore likely to need some work. Possibly, the items in the Writing Style Questionnaire which deal solely with the amount of revision need reconsideration, as these might not be central to the definition of a reviser writing style. An example of such an item would be this statement: "When my text is ready, I elaborately read through it and make improvements: a lot can still be changed at that point". A Writing Style Questionnaire item which typically represents the part of the definition focusing on using text production to construe a content plan is: "For me, writing is a way to get my thoughts clear". On the basis of the present results, then, it seems than the planner dimension of the Writing Style Questionnaire can better predict different online configurations than the reviser dimension.

Cromley and Azevedo (2006) and Veenman et al. (2003) found that offline reports had little or no predictive value for online task execution. In the present study, relations between self reports and online task execution have been established, as is the case in the study by Torrance et al. (1999). There are two main differences between these two sets of studies which may explain the different findings. The first is that in the present study and the study by Torrance et al. (1999), the data are analyzed temporally, which is not the case in the studies by Cromley and Azevedo (2006) and Veenman et al. (2003). However, the absence of a temporal analysis in the former studies cannot be the sole explanation for the absence of relations between offline and online data. First of all, the fact that frequential main effects were found in the present study for text production, reading the assignment and revising, demonstrates that a temporal analysis is not always needed. In addition, Cromley and Azevedo (2006) and Veenman et al. (2003) studied reading tasks, whereas the present study and the study by Torrance et al. (1999) deal with writing processes. The reported writing styles – planner and reviser styles – imply variation in distributions, whereas this is not so much the case for the offline measures used by Cromley and Azevedo (2006) and Veenman et al. (2003). Whereas there is evidence available that the occurrence of cognitive activities can also vary across the reading process (Janssen, Braaksma, Rijlaarsdam & Van den Bergh, 2005), this is not what offline measures of reading tasks generally focus on. It follows that the predictive value of these particular offline measures cannot be analyzed temporally.

In this study the planning activity was construed of five subcategories (monitoring, goal setting, generating, structuring, metacomments). In future research, however, the validity of the analysis could possibly be increased by modeling the occurrence of these subcomponents separately. Hayes & Nash (1996), for example, distinguish between 'content planning' and 'non-content planning'. Goal setting, generating and structuring might arguably be instances of content planning, whereas monitoring and metacomments are more process-oriented activities and might therefore be seen as instances of non-content planning. Ideally, the relation between reported planner style and online planning activities should be analyzed separately for different types of online planning. This was not possible in the present study, due to the low frequency with which the subcomponents occur.

To conclude, it seems that questionnaires can have predictive value for online task behavior. Kieft's (2006; 2008) Writing Style Questionnaire, and particularly its planner scale, seem to be a valid predictor of writing processes. In addition, a temporal analysis of (meta)cognitive activities across task execution seems to be a valid and sensitive reflection of online processing, particularly for writing. Whether a temporal analysis is also suitable for bringing out relations between offline and online measurements for other types of tasks, such as reading and mathematic tasks, is an issue for future research.

CHAPTER 6

CONCLUSION AND DISCUSSION

The main aim of the research reported in this thesis was to explain quality differences between L1 and L2 writing. To do so, data on L1 and L2 writing were collected among twenty fourteen- and fifteen-year-old students of secondary education. Their L1 was Dutch. This study's L2 was English. On average, students had followed English as a school subject for four or five years, for approximately two or three hours a week. In addition, English is frequently used in Dutch media. Each student wrote four short argumentative essays in L1, and four short argumentative essays in L2. Their writing processes were registered by means of think aloud procedures, combined with keystroke logging. Analysis of the writing processes involved the following cognitive activities: reading the assignment, (process and content) planning, formulating, reading own text, evaluating own text, and revising. The students also completed language proficiency tests in both Dutch (L1) and English (L2), and a questionnaire in which they reported on their writing style.

DISCUSSION AND INTERPRETATION OF THE MAIN FINDINGS

While it is the consensus that students' L2 texts are generally of lower quality than their L1 texts, it had not yet been possible to quantify this quality difference. Researchers had, in previous research, compared isolated features of L1 and L2 writing. L2 texts were, for example, found to be shorter, to contain more linguistic errors, to contain less cohesive argumentation, and to be less focused on the reader (Silva, 1993). However, L1 and L2 text scores which express students' writing proficiency as a whole could generally not be compared. For this to be possible, the quality of both L1 and L2 texts must be expressed on the same scale. This is normally not the case, for various reasons, such as raters' different attitudes towards L1 and L2 writing, which causes them to be more strict or lenient to one of the two languages.

In **chapter 2** of this thesis, a procedure was presented and tested which makes direct comparisons of L1 and L2 text scores possible. Two main features define this procedure: 1) raters are bilingual or near native users of both L1 and L2, which increases the chance that they are equally strict or lenient in rating L1 and L2 texts; 2) ratings are performed with L1 and L2 benchmark texts (i.e. texts representing average quality). All texts were rated on three criteria of text quality: global quality, structure, and language. The results showed that the ratings with L1 and L2 benchmarks were parallel tests and that the ratings were performed reliably. These outcomes show that the raters did not apply rating standards differently in the L1 and L2 condition. Direct comparisons of observed L1 and L2 text scores were therefore warranted. In the investigated population, L2 text scores are much lower than L1 text scores. Effect sizes (*Cohen's d*) indicate, for example, that the difference between L1 and L2 global quality text scores is the size of approximately five standard deviations of the score differences due to tasks. While different tasks (within a language) were found to cause large score differences, the score difference due to having to write in a different language is even larger. This means that the L1 and L2 texts are largely separate samples, and that there is only a very small overlap area where the highest scoring English essays and the lowest scoring Dutch essays meet.

Since L1 and L2 texts were now rated on the same scale, it was also possible to make a direct comparison of L1 and L2 relations between writing processes and text quality. This was done in **chapter 3**. Such a comparison allows us to investigate whether effective L2 writing processes are different from effective L1 writing processes. In other words: to write L2 texts of as high a quality as possible, should writers perform their L2 writing processes as they did in L1, or should their L2 writing processes be different, to accommodate the added difficulty of L2 writing (cf. van Weijen, 2009)?

Writing processes are characterized by specific configurations of their constituent cognitive activities, such as generating content, planning, formulating (text production), structuring and making revisions (Flower & Hayes, 1980; Hayes & Flower, 1980; Hayes, 1996). These cognitive activities can (re)occur at (more or less) any given moment during the writing process. Indeed, it has been shown that the quality of writing process execution is reflected by the temporal distribution of cognitive activities across task execution: it matters at which moment during the writing process specific cognitive activities occur (Breetvelt et al., 1994; Leijten & Van Waes, 2006; Olive et al., 2008; Roca de Larios et al., 2001; Rijlaarsdam & Van den Bergh, 1996; Rijlaarsdam et al., 2011; Van den Bergh & Rijlaarsdam, 1996; Van den Bergh et al., 2009; Van der Hoeven, 1997). For example, Rijlaarsdam and Van den Bergh (1996) demonstrated for L1 writing that structuring activities were more effective when they occurred during early stages of task execution (i.e. the correlation between structuring and text quality is at its highest at the start of the writing process) and less effective when they occurred towards the end of task execution.

In chapter 3, writing processes, and relations between writing processes and text quality, were therefore analyzed in terms of the temporal distributions of cognitive activities. The temporal dimension of the model, i.e. 'time during the writing process', was operationalized by splitting each protocol into five episodes, each episode containing twenty percent of the segments in that think aloud protocol. Episode 1 represented the start of task execution, whereas episode 5 reflects the end of task execution, for example. The results showed that L1 and L2 writing processes were quite similar (in terms of temporal distributions of cognitive activities), but that cognitive activities contribute to text quality differently during L1 and L2 writing. Two main differences were found. First, cognitive activities are relevant to text quality at different stages of task execution during L1 and L2 writing. For example: content planning is related to text quality at the start and during the middle part of the writing process in L1, but only at the end of task execution in L2. This means that, for L1 and L2 writing, there are different crucial moments during task execution, at which it matters whether writers apply a specific cognitive activity more or less often. In other words, writers need to distribute their attention differently across L1 and L2 writing tasks. Effective L2 writing processes are different from effective L1 writing processes.

Second, where a cognitive activity is relevant to text quality in the same stages of task execution during L1 and L2 writing, the activity is most often positively related to text quality in L1, but negatively in L2. For example: reading the assignment is positively related to text quality just after the start of task execution in L1, but negatively in L2. And: evaluating the text-written-so-far is positively related to text quality at the end of task execution in L1, but negatively in L2. How can it be the case that performing the same writing process activity at the same moment (e.g. evaluating your text at the end of task execution) is a feature of successful L1 writing, but a feature of less proficient writing in L2?
Although this question could not be answered based on the data and results of chapter 3, it seems plausible that the negative effects of cognitive activities on L2 text quality are related to language proficiency problems, which are, in the investigated population, much more present during L2 writing than during L1 writing. Indeed, it has been hypothesized that language difficulties constrict working memory resources during L2 writing, leaving fewer resources to perform conceptual and regulatory activities (such as evaluating) in a sufficient manner (Sasaki & Hirose, 1996; Schoonen et al., 2003). Possibly, writers were able to engage in cognitive activities at 'the right moment' during L2 writing (that is, 'the right moment' according to L1 standards, e.g. performing evaluations at the end of task execution), but these activities might have been lacking on different aspects of quality, due to constricted working memory resources, or simply due to insufficient

knowledge of (text conventions in) the English language. An example of such an additional quality aspect of cognitive activities, on which L2 writing processes might have been lacking, involves the 'objects' of cognitive activities. For example: evaluations which are constricted to local concerns of the produced text, such as the spelling of words and the grammar of sentences, are qualitatively different from evaluations pertaining to the rhetorical structure of the entire text. Second, the quality with which cognitive activities are executed may be expressed in terms of their interrelations with other cognitive activities, which precede or follow it. Van den Bergh and Rijlaarsdam (1999) showed that the function of a cognitive activity can be expressed by investigating what kind of activity precedes it. They distinguish different kinds of 'content generating', such as 'Assignment-Driven-Generation' (new ideas are sparked by reading information in the assignment) and 'Translation-Driven-Generation' (new ideas are sparked as the writer formulates text; the activity of putting thoughts into language can be an incentive for generating new content). These two types of 'content generating' are qualitatively different. Similarly, the quality of evaluations of the produced text might be characterized by whether or not it is followed by a (successful) revision.

If this line of reasoning is adopted, it follows that relations between cognitive activities and text quality are expected to be different for different levels of language proficiency. To explore this hypothesis, language proficiency scores were added to the regression model used in chapter 3, as *z scores* ('*zLP*'). The language proficiency variable concerned L1 language proficiency scores for writing tasks carried out in L1, and L2 language proficiency scores for writing tasks carried out in L2. The result was a model to explain text quality, in which the predictor variables are: *language proficiency scores*, *episode 1*, *episode 2*, *episode 3*, *episode 4*, *episode 5*, and interaction variables *language proficiency*episode 1*, *language proficiency*episode 2*, *language proficiency*episode 3*, *language proficiency*episode 4*, *language proficiency*episode 5*. This analysis was carried out for each of the cognitive activities included in this thesis. The (significance of the) regression weights for the interaction variables indicate whether and how the effect of a cognitive activity, per episode, on text quality is influenced by language proficiency. The exact regression weights are reported in Appendix N. Table 1 gives an overview of the established interaction effects, for both L1 and L2 writing. '+' reflects a positive regression weight, '-' indicates a negative regression weight. Remember that the language proficiency scores were entered as *z scores*, so that lower language proficiency scores had negative values, and higher language proficiency scores had positive values. Therefore, pluses and minuses in table 1 should be interpreted as follows:

- : cognitive activity is negatively related to text quality for students with higher language proficiency scores (*zLP > 0*), but positively related to text quality for students with lower language proficiency scores (*zLP < 0*)

+ : cognitive activity is positively related to text quality for students with higher language proficiency scores (*zLP > 0*), but negatively related to text quality for students with lower language proficiency scores (*zLP < 0*)

Table 1 shows that, where cognitive activities are relevant during the same stages of task execution (i.e. reading the assignment in episode 2, process planning in episodes 2 and 3, revising in episodes 3 and 4), the direction of the relationship (i.e. + or -) is similar for L1 and L2 writing, if language proficiency is considered as a mediating predictor. This is contrary to the outcomes of chapter 3, where text

*Table 1. Interaction effects of language proficiency *cognitive activities on text quality per episode*
L1 = effect of activity for L1 writing; L2 = effect of activity for L2 writing
Empty boxes indicate that no significant effects could be established (p > .05)
zLP = language proficiency (z score)

		Epi1*zLP	**Epi2*zLP**	**Epi3*zLP**	**Epi4*zLP**	**Epi5*zLP**
Reading Assignment	L1		-		-	
	L2		-		-	
Process Planning	L1		+	-		
	L2		+	-		-
Content Planning	L1	-				
	L2					
Formulating	L1					
	L2		-	+		
Reading Own Text	L1				+	
	L2					-
Evaluating Own Text	L1					
	L2					
Revising	L1			-	+	
	L2			-	+	

quality was predicted by episodes (i.e. 'time during the writing process') only, and language proficiency was not included in the model. In chapter 3, cognitive activities which were relevant in the same episodes, were mostly positively related to text quality during L1 writing, but negatively during L2 writing.

It should be noted that some caution is warranted in interpreting the directions (+ and -) reported in table 1. Unlike the text quality scores, the L1 and L2 language proficiency measurements are probably not parallel tests (although this was not tested), so that the L1 and L2 regression weights, and therefore pluses and minuses in table 1, may strictly speaking not be assumed to be directly comparable. But, although strong conclusions are not warranted, the outcomes presented in table 1 seem to suggest that, if language proficiency is taken into account, L1 and L2 writing are quite similar in their demands in terms of how often activities are applied during certain episodes of task execution. The outcomes in table 1 also suggest that, for higher levels of language proficiency, cognitive activities are executed with similar quality in L1 and L2. Future research could further investigate what this quality entails. Is the object of cognitive activities, for example, a valid operationalization of process quality, in addition to the temporal approach? In addition, future research may investigate this issue more thoroughly by creating parallel measurements of L1 and L2 writing proficiency.

In **chapter 4**, the effect of (L1 or L2) language proficiency on task execution during L1 and L2 writing was investigated. Here, too, writing process execution was analyzed in terms of temporal distributions of cognitive activities. These distributions differ between writers. The main aim of the research presented in chapter 4 was to find out whether these inter-individual differences could be explained by differences in language proficiency. This was done for L1 and L2 writing separately. It was expected that language proficiency influences the distributions of strongly linguistic activities, such as formulating, but also the distributions of more conceptual and regulatory activities, such as planning. However, effects were only found on evaluating own text in L1 and on process planning in L2. During L1 writing, students with higher (L1) language proficiency scores are more likely than average to evaluate their texts. Evaluation differences between students due to (L1) language proficiency are equally large throughout the L1 writing process. L2 process planning is more likely to be carried out by students with higher (L2) language proficiency scores throughout the writing process, but the differences due to language proficiency are largest at the start of the writing process. That is, the occurrence of process planning varies more across task execution for students with higher (L2) language proficiency: the bulk of process planning activities is carried out at the start of the writing process. Students with lower (L2) language proficiency showed less variation in process planning across the writing process: process planning activities are equally likely to occur at the start and end of task execution, although there is drop in the occurrence of process planning in episode 3.

At this point, it interesting to combine these results with the obtained knowledge about when (during the writing process) cognitive activities are most beneficial to text quality for more and less language proficient students (cf. table 1 of the current chapter). In episode 2, L2 process planning is positively related to text quality for students with higher language proficiency scores, but negatively for students with lower language proficiency scores. In episodes 3 and 5, L2 process planning is negatively related to text quality for students with higher language proficiency scores, but positively for students with lower language proficiency scores. In other words, students with higher (L2) language proficiency write better texts if they perform more process planning at the start (episode 2) of task execution, and less process planning during later stages of the writing process (episodes 3 and 5). In other words, students with higher (L2) language proficiency distribute their process planning activities in an effective manner, on average. Students with lower (L2) language proficiency, on the other hand, write better texts if they perform less process planning at the start (episode 2) of task execution, and more process planning during episodes 3 and 5. This routine is only partly followed by students with lower language proficiency. They plan less than average in episode 2, and show an increase of process planning in episode 5, which is effective in terms of text quality. However, they plan hardly at all in episode 3, even though, for them, process planning is positively related to text quality in this episode.

Interestingly, no effect of language proficiency on formulating – the most language-specific cognitive activity included in the analysis – could be established in chapter 4. This might be explained by two methodological features of the research reported in chapter 4. First, the participants in the study – fourteen- and fifteen-year-old students of pre-academic secondary education – might all of them have reached such a level of language proficiency that any inter-individual language proficiency differences are too small to be a substantial influence on the quality (i.e. temporal distribution) of formulating during writing. However, and this is the second methodological feature to discuss, temporal distributions may reflect a substantial part of the quality of writing process execution, but possibly not all of it, as was also noted in the discussion of chapter 3 on pages 55-56. Even if all students distribute their cognitive activities similarly across the writing process, the activities may still differ on other aspects of processing quality. Two such aspects were already mentioned, namely the ‘objects’ of activities, and interrelations with other cognitive activities (cf. p. 78). So while temporal distributions of cognitive activities have been shown to be a valid reflection of the quality of the writing process, the quality of task execution may additionally be reflected in terms of the 'objects' of activities, and interrelations with other cognitive activities.

The focus of **chapter 5** is somewhat different from that of previous chapters. It deals with L1 writing only, and is less focused on uncovering the writer or process characteristics responsible for successful writing. Rather, chapter 5 has a methodological focus. It tests the validity of a questionnaire (Kieft et al., 2006, 2008) in which respondents report their writing style, by investigating the questionnaire's predictive value for actual writing behavior during task execution. A number of researchers have, on the basis of their research, suggested that many writers' writing styles can be typified by the degree to which they are *planners* or *revisers* (Biggs et al., 1999; Kieft et al., 2006, 2008; Torrance et al., 1994, 1999, 2000). Typical planners are assumed to make an extensive (mental) content plan before commencing with text production. Revisers, on the other hand, are assumed to use text production as a means to arrive at a content plan. Revisers need more revisions, as their initial pieces of text are written before their content plans are complete. It was therefore expected that higher or lower degrees of reported planner style or reviser style entail different temporal distributions of cognitive activities. This hypothesis was partly confirmed. Different distributions due to reported writing style were found for reading the assignment and planning. Students with higher reported reviser scores are (slightly) less likely than average to read the assignment at the start of task execution and (slightly) more likely than average to read the assignment at the end of the writing process. Students with higher reported planner scores are more likely than average to plan at the start of the writing process, but less likely to plan during the remainder of the writing process. In addition, a number of main effects of planner and reviser scores were established. For example: the higher the reported planner score, the more likely it is that a student engages in formulating and revising throughout the writing process. These findings demonstrate that Kieft's (2006; 2008) questionnaire has predictive value for the writing process, although a number of problems were signaled, too. It is, for example, rather striking that the reviser dimension in the questionnaire had no predictive value for the occurrence of formulating and revising during writing. It was recommended that the reviser dimension in the questionnaire should be worked on, and tested, before its outcomes are used as a variable in future studies.

METHODOLOGICAL CONSIDERATIONS

Rijlaarsdam and Van den Bergh (1996) and Van den Bergh and Rijlaarsdam (1996) introduced the idea that the function of a cognitive activity depends on the context in which it occurs. For example: reading the assignment at the start of task execution probably serves a different function than reading the assignment during final stages of the writing process. At the start of task execution, reading the assignment is a means of determining the intended text's topic, goals, audience, et cetera. Towards the end of task execution, reading the assignment is more probably an aspect of evaluative activities: does the produced text satisfy the assignment's requirements? Researchers who want to investigate the quality of cognitive activities during writing therefore need to take the context in which the activities occur into account.

Temporal analyses of writing processes

One way to achieve this is by adopting a temporal approach to writing processes: researchers should include 'time during the writing process' as a proxy variable for the context in which cognitive activities are applied (Rijlaarsdam & Van den Bergh, 1996; Van den Bergh & Rijlaarsdam, 1996). When researchers analyze writing processes temporally, they face at least two methodological questions, the answers to which imply a certain theoretical viewpoint. These two methodological issues are discussed below.

Operationalizing 'time during the writing process'

In the research presented in this thesis, 'time during the writing process' was approximated by using segments which start whenever a new cognitive activity starts. Alternatively, 'time during the writing process' can be measured in minutes, seconds, and/or milliseconds. Both methods have their respective drawbacks and strong points.

The drawback of measurement in terms of segments, is that the weight of each segment is equal in the analysis, regardless of how much content the segment contains. For example, qualitative exploration of the think aloud protocols suggests that, in some protocols, formulating activities are often interrupted by new thoughts ('content planning'), or different cognitive activities, while in other protocols this seems to happen less. So, it may easily be the case that, even if equal amounts of content are formulated in protocols A and B, this content is spread out over more segments in protocol A, so that protocol A consists of more segments containing formulating activities than protocol B. The problem of measurement in terms of segments, then, is that each segment is weighted equally in the analysis, even if they may reflect unequal amounts of content.

However, time registration in minutes and seconds also has a weighting drawback, which is the reverse of the weighting problem with segments. Equal amounts of attention paid to a cognitive activity are not always reflected by equal amounts of time (in, for example, seconds) spent on an activity, resulting in unequal weighting in the analysis. Inherent to measurement in minutes and seconds is that the weight or importance of an activity during task execution is expressed in how much time it takes up. This presupposes that the quality with which an activity is carried out is reflected by how much time a writer spends on this activity. However, the amount of time spent on an activity does not necessarily reflect the amount of attention paid to it.

If it is the object of the researcher to measure the amount of attention paid to – or cognitive effort put into – cognitive activities by writers, measurement in terms of segments is probably quite valid. After all, if writer A's writing process contains more segments with revision activities than writer B's writing process, this means that writer A re-involved him- or herself in revising more often than writer B, or, in other words: that writer A renewed his or her attention to revising more often than writer B. It may therefore be argued that measuring 'time during the writing process' in terms of segments was, for the research reported in this thesis, a suitable choice of method.

In any case, the choice for measurement in terms of segments or minutes and seconds should probably be tuned to the research question. Researchers should decide which method of measurement is a better conceptualization of writing process quality for their specific research goals.

Standardization

In the research reported in this thesis, the length of writing processes was standardized by splitting each writing process into five episodes, and calculating, for each cognitive activity, its percentage (or proportion) of occurrence per episode relative of the total number of segments in that episode. A protocol of 330 segments, for example, would be analyzed as five episodes consisting of 66 segments each. Similarly, a protocol of 125 segments would be analyzed as five episodes consisting of 25 segments each. This way, each episode 1 (or 2, or 3, et cetera) is directly comparable with episode 1 (or 2, or 3, et cetera) in another protocol. Standardization, then, cancels out differences in length (in this thesis: differences in number of segments) between writing processes.

The choice for or against standardization is, like the choice of how to measure 'time during the writing process', pre-eminently a theoretical one. It depends on the relevance which is attached to differences in length (in segments or minutes) of writing processes.

An objection to standardization (by means of episodes) is a loss of statistical power: a writing process with – for example – 600 segments, or observations, is reduced to a writing process with five observations: one proportion per episode. The price that researchers pay for such a form of standardization is that effects are underestimated. This is advantageous on the one hand: estimations are conservative, and false positives are therefore unlikely. On the other hand, this means that the more subtle effects are more easily overlooked. Another objection to using standardization is that differences in length of writing processes are likely to reflect qualitative differences in the execution of the writing process. If a writer needs double the amount of segments, this says something about this writer, or about this writing process. Segment 600 out of a 600-segment-protocol (or the 27th minute of a 27-minute-protocol) is conceptually different from segment 300 out of a 300-segment-protocol: more cognitive activities have preceded it.

However, the reverse argumentation may be used in favour of standardization. Without standardization, segment 300 out of a 600-segment-protocol is compared to segment 300 out of a 300-segment-protocol, whereas they are likely to differ conceptually. It is comparing middle parts with end parts of task execution. In addition, standardization has the benefit of more reliable estimates of average occurrences of cognitive activities during final stages of process execution, as these estimates are based on the complete test sample, instead of on those participants who have not yet finished writing. Estimations of average distributions of cognitive activities with non-standardized writing processes, then, become more and more unreliable towards the end of task execution (cf. Van den Bergh & Rijlaarsdam, 2001).

Interactions between cognitive activities

In the reported research, the temporal distributions of the various cognitive activities were analyzed separately. That is, the distribution of reading the assignment was modeled independently of the distribution of process planning, which was in turn modeled independently of the distribution of content planning, and so forth. In reality, of course, cognitive activities interact. For example: if process planning does not occur until episode 3 – which is fairly late in the writing process – this is likely to be an artifact of problems with other cognitive activities during episodes 1 and 2. The writer may, for example, have had trouble understanding the assignment, and may therefore have needed to continue reading the assignment through episode 2. Indeed, that process planning is mostly negatively related to text quality during the final stages of the writing process (cf.

chapter 5, table 6; this chapter, table 1) might not in the first place be because of the nature of process planning, but rather because process planning is taking the place of some activity which is more relevant at the end of task execution (e.g. reading own text, or revising). Van den Bergh and Rijlaarsdam (1999) also show that the likelihood that cognitive activity Z occurs can be predicted from whether cognitive activity Y occurs first. For example: the occurrence of formulating increases the probability that generating occurs.

The interactions between cognitive activities can be taken into account by using multivariate analyses. Such analyses are another way to model the context in which cognitive activities occur, in addition to the temporal approach of writing processes. Applying multivariate analyses was, however, not feasible in the research reported in this thesis, due to the moderate sample size.

SUGGESTIONS FOR FUTURE RESEARCH

On the basis of the research reported in this book, a number of suggestions for future research of writing can be made.

Multiple tasks

In chapter 2, it was demonstrated that there was large task-related variance in the text scores, even though tasks only differed in terms of topic. This indicates that measurements with few assignments are unreliable representations of writing proficiency. It was estimated, for example (see chapter 2), that a measurement of global quality based on one assignment (=topic) has a reliability of between .33 and .39. The reliability of the measurement is .80 to .83 if eight tasks are used per writer.

This reinforces a point made by many researchers before, namely that measurements of writing skill should involve multiple tasks per writer (Coffman, 1966; Gebril, 2009; Schoonen, 2005; Van den Bergh, 1988b; Wesdorp, 1974). Indeed, Van den Bergh (1988b) argues that writing assessments on the basis of single tasks are basically single-item-tests, which do not allow for generalizations about an individual's writing proficiency. The minimum number of tasks to be used for a reliable assessment of writing skill depends, among other things, on the applied rating method, number of raters, test population, and task type (cf. Coffman, 1966; Schoonen, 2005; Van den Bergh, 1988b).

Text quality

Independence of rating criteria and halo effects

Also in chapter 2, strong (disattenuated) correlations were found between the three rating criteria used: global quality, structure, and language. To some extent, this reflects a true coincidence of rating criteria (cf. Bae and Bachman, 2010). Global quality includes the two other rating criteria, structure and language. The strong correlations between global quality and structure and between global quality and language are therefore not automatically indicators of major validity problems with the applied rating criteria. After all, if criteria (partly) coincide, they are expected to correlate. However, the perfect correlation between global quality and structure probably exceeds the expected correlation and indeed raises questions about the validity of these two rating criteria. Similarly, the correlation ($r = .75$) between structure and language, two rating criteria which are assumed to encompass distinct aspects of text quality, is probably higher than expected, too. It suggests that raters were suffering from halo effects, even though care was taken in the present study to minimize this possibility.

In future research, it might be interesting to investigate whether such strong correlations between rating criteria are also found if the ratings of different criteria are carried out by different rater teams (who should among themselves be comparable, of course). If so, this would imply that the strong correlations are probably not the result of a halo effect, but that the distinction between rating criteria might be artificial.

Holistic versus analytic ratings

The ratings conducted in chapter 2 are holistic ratings: each text is awarded one score (per rating criterion), which is assumed to reflect the entire quality of the text (for that specific rating criterion). This is in contrast to analytic ratings, where each text is awarded multiple score, for several specified aspects of writing (Weigle, 2002). Holistic ratings have been claimed to be more valid than analytic ratings (cf. Schoonen, 2005, who demonstrates that holistic ratings ("collected with essay scales") have higher generalizability than analytic scores ("with scoring guides"); White, 1984, 1985, as cited by Weigle, 2002). One of the main validity-related advantages of holistic scoring is that it may be more sensitive to capturing the reaction of the reader (White, 1984), which is of course an essential feature of the quality of the communicative act that writing is. Analytic ratings, on the other hand, are sometimes claimed to have a reliability advantage due to multiple scores being assigned per text (for discussions of holistic versus analytic rating, see Rijlaarsdam et al. 2011; Weigle, 2002, p. 72-73 and p. 112-121).

Nevertheless, the jury reliabilities for the holistic ratings in chapter 2 were quite satisfactory. So, as holistic ratings are probably more valid, as they are more efficient (i.e. faster), and as they have been shown to be quite reliable – if carried out by multiple raters and with the aid of benchmark essays – it seems that holistic ratings are a good method for establishing text quality.

Analyzing the quality of writing process execution

In this book, writing processes, and relations between writing processes and text quality, were analyzed temporally. The moments at which cognitive activities (such as planning, formulating, and revising) occur during writing (i.e. at the start or toward the end of task execution) reflect the quality of the writing process, rather than how often cognitive activities occur. This is conform earlier research (Breetvelt et al., 1994; Leijten & Van Waes, 2006; Olive et al., 2008; Roca de Larios et al., 2011; Rijlaarsdam & Van den Bergh, 1996; Rijlaarsdam et al., 2011; Van den Bergh & Rijlaarsdam, 1996; Van den Bergh et al., 2009; Van der Hoeven, 1997).

The research reported in this thesis confirms that the temporal approach is a valid reflection of the quality with which the writing process is executed. In chapter 3, for example, regression models in which text quality scores were explained with percentages of cognitive activities per episode, were better fits to the data than models in which text quality scores were explained by the average occurrence of cognitive activities across the whole writing process. In addition, the moderately satisfactory (cf. chapter 3, table 4) and satisfactory (cf. chapter 4, figure 1) correlations between observed writing process data and the values as predicted by temporal models also indicate that the temporal approach captures a substantial part of (the observed) writing process quality.

Nevertheless, the finding that cognitive activities can, in the same episode, be positively related to text quality in L1, but negatively in L2 (cf. chapter 3), indicates that the moment at which a cognitive activity is applied does not capture the quality with which is executed entirely. Therefore, it might be useful, as was suggested in chapters 3 and 4, if researchers take into account additional features of processing quality, besides the temporal approach. Earlier in chapter 6 (cf. table 1), it was demonstrated that these additional features of quality are likely to be explained by language proficiency. Additional features of writing process quality are probably different for different cognitive activities, but it seems worthwhile to look into features such as the objects of cognitive activities (e.g. are evaluations of the text constricted to local concerns, such as the spelling of words and the grammar of sentences, or do the evaluations pertain to the rhetorical structure of the entire text? cf. Stevenson et al., 2006) or the interrelations between cognitive activities (e.g. reading the assignment probably has a different function when it is followed by the generation of new ideas than when it is followed by revisions of the text produced so far).

Another thing to note about the temporal approach is that the temporal analyses can be carried out in different manners. In chapter 3, for example, the occurrence of cognitive activities was estimated for each episode separately (and the difference between episodes tested). This analysis, then, involves investigations *within* episodes. In chapter 4, on the other hand, 'episode' was included as a scale (predictor) variable, and transitions in the occurrence of cognitive activities *between* episodes were also modeled. Another difference between the temporal analyses in chapter 3 and 4 is that chapter 4's model is a polynomial: it can include non-linear terms (i.e. powers of episode: episode^2, episode^2, et cetera), whereas chapter 3's model cannot. In short, the model in chapter 4 is more fine-grained. It is therefore not surprising that the correlations between predicted and observed values were stronger in chapter 4 than in chapter 3 − although the strength of the correlations was already (moderately) satisfactory in chapter 3. If feasible, then, polynomial functions in which the occurrence of cognitive activities is described as a function of episode are probably to be preferred for analyzing writing processes temporally. (In chapter 3, the 'within-episodes' model was a more suitable choice, as it was more in line with the second type of analysis in the chapter, in which relations between cognitive activities and text quality were estimated per episode. The latter was also done by means of a within-episode model, so as to keep the number of parameters limited: because of the direct comparison between process-product relations in L1 and L2, the model was already very complex.)

Mediating factors

In this thesis, the influence of two writer variables on task execution has been investigated: language proficiency in chapter 4, and writing style (i.e. planner and reviser styles) in chapter 5. Both variables were found to be predictive of the manner in which writing processes are carried out, although effects were sometimes small. In addition, it was demonstrated earlier in chapter 6 that language proficiency seems to mediate the relations between writing process and text quality. That is, whether applying cognitive activities (during specific stages of task execution) is beneficial to text quality may depend on the writer's level of language proficiency. For example: revising during pre-final stages of the writing process (i.e. in episode 4) becomes more beneficial to text quality as language proficiency increases.

It could increase our understanding of the factors underlying successful writing if future research further investigates the mediating role of language proficiency on relations between cognitive activities during writing and text quality. A relevant question could be: in what way does language proficiency affect the quality with which cognitive activities are carried out? And: is it possible to identify a threshold level of language proficiency beyond which the quality of writing becomes less dependent on language proficiency?

In addition, future research might look into the mediating role of writing style on relations between writing process and text quality. The degree to which writers are planners or revisers might, like language proficiency, mediate relations between the occurrence of cognitive activities during writing and text quality. For example, it may be expected that (content) planning at the start of task execution is less effective for typical revisers.

Replication

The participants in this thesis constituted a fairly homogeneous sample. They were all, for example, fourteen- and fifteen-year-old students of pre-academic secondary education. Results might be different for different types of learners. Van der Hoeven (1997), for example, investigated the role of linguistic proficiency on writing processes among twelve-year-old students and found that linguistic proficiency affected the occurrence of structuring activities, but also the temporal distributions of generating and formulating (Van der Hoeven, 1997, p. 112-115). So here, the execution of formulating – a strongly linguistic cognitive activity – is affected by linguistic proficiency, whereas this is not the case in the research reported in this thesis (cf. chapter 4). Another difference between Van der Hoeven's (1997) findings among twelve-year-olds and the findings among fourteen-and-fifteen-year-olds reported in this thesis, concerns the moments during writing at which cognitive activities are most strongly related to text quality. For example, in Van der Hoeven's study, reading own text is most strongly related to text quality during the centre part of task execution. From thereon until the end of task execution, the effectiveness of reading own text (for text quality) decreases. In chapter 3 of this thesis, however, reading own text is most strongly related to text quality during final stages of task execution. Possibly, the described differences are due to the developmental stage of the writers. Ideally, a longitudinal study should be set up to investigate what 'successful writing' means at different developmental stages.

Whereas Van der Hoeven (1997) investigated writing proficiency among students younger than the ones in this thesis, Van Weijen (2009) investigated writing skill among first-year-university students. While there are quite a number of similarities between Van Weijen's (2009) findings and the findings presented in this book, there are some striking differences, too. One of these differences is that the variation between writing processes due to task seems to be much larger in Van Weijen's (2009) study than in the present study, where hardly any task-related variance between writing processes was found (cf. Appendix I). Future research could further investigate whether task-related process variation is more common among more advanced writers.

Other second languages

In the present study, the L2 was English, and the L1 was Dutch. The investigated L1 and L2 were therefore languages from cultures which are relatively similar, which means that ideas about what constitutes a good piece of writing are not likely to differ greatly. The added difficulty of L2 writing is in this case therefore mostly made up of the extra cognitive burden of having to write in a non-native language. In addition to the 'normal' cognitive burden of writing, extra restrictions are placed upon working memory by language difficulties.

If L1 and L2 are from cultures which are less similar, genre-related difficulties are likely to be added to the cognitive load, besides language difficulties. For example, ideas about what kind of argumentation (e.g. deductive vs. inductive reasoning) is persuasive are largely culturally bound. Apart from affecting the idea of what constitutes good writing, i.e. the conceptualization of text quality, cultural differences are also likely to affect the writer's knowledge-base in long-term memory (cf. Hayes, 1996). After all: if the L2 is from a culture very dissimilar to that of the L1, the writer's L2 genre knowledge (but also his or her audience knowledge, for example) is likely to be smaller during L2 writing than during L1 writing. This will probably influence this writer's writing process. Therefore, it would be useful to investigate whether the L1/L2 differences reported in this study are generalizable to other combinations of first and second languages.

REFERENCES

Bae, J., & Bachman, L.F. (2010). An investigation of four writing traits and two tasks across two languages. *Language Testing*, *27* (2), 213-234.

Biggs, J., Lai, P., Tang, C., & Lavelle, E. (1999). Teaching writing to ESL graduate students. A model and an illustration. *British Journal of Educational Psychology*, *69*, 293-306.

Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement*, *22*, 41–52.

Breetvelt, I., Rijlaarsdam, G., & Van den Bergh, H. (1994). Relations between writing processes and text quality. When and how? *Cognition & Instruction*, *12* (2), 103-124.

Chenoweth, N.A., & Hayes, J.R. (2001). Fluency in writing. Generating text in L1 and L2. *Written Communication*, *18* (1), 80-98.

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

Couzijn, M., Van den Bergh, H., & Rijlaarsdam, G. (2002). Writing in L1 and L2. Does it make a difference? Staffordshire: Presentation at EARLI SIG Writing Conference, 11-13th July 2002.

Cromley, J.G., & Azevedo, R. (2006). Self-report of reading comprehension strategies. What are we measuring? *Metacognition and Learning*, *1*, 229-247.

De Glopper, K. (1988). *Schrijven beschreven. [Writing described.]* 's-Gravenhage, the Netherlands: SVO.

Flavell, J.H. (1979). Metacognition and cognitive monitoring. A new area of cognitive-developmental inquiry. *American Psychologist*, *34* (10), 906-911.

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing. Making plans and juggling constraints. In L. W. Gregg, & E.R. Steinberg (Eds.). *Cognitive processes in writing* (pp. 31-50). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Galbraith, D. (1996). Self-monitoring, discovery through writing, and individual differences in drafting strategy. In G. Rijlaarsdam, H. van den Bergh, & M. Couzijn (Eds.). *Theories, models and methodology in writing research* (pp. 121-141). Amsterdam: University Press.

Galbraith, D., & Torrance, M. (2004). Revision in the context of different drafting strategies. In G. Rijlaarsdam (Series Ed.), L. Allal, L. Chanquoy, & P. Largy (Vol. Eds.). *Studies in Writing. Vol. 13. Revision: Cognitive and instructional processes* (pp. 63-85). Dordrecht: Kluwer.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one method fit it all? *Language Testing*, *26,* 507-531.

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

Goldstein, H. (1979). *The design and analysis of longitudinal studies. Their role in measurement of change.* London: Academic Press.

Gorman, T.P., Purves, A.C., & Degenhart, R.E. (1988). *The IEA study for written composition I : the international writing tasks and scoring scales.* Oxford: Pergamon Press.

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

Hayes, J. R., & Flower, L. S. (1980). Identifying the organization of writing processes. In L. W. Gregg, & E.R. Steinberg (Eds.). *Cognitive processes in writing* (pp. 3-30). Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hayes, J.R., & Nash, J.G. (1996). On the nature of planning in writing. In C.M. Levy, & S.Ransdell (Eds.). *The science of writing. Theories, methods, individual differences, and applications* (pp. 29-55). Mahwah, N.J.: Lawrence Erlbaum Associates.

Healy, M.J.R. (1989). Growth curves and growth standards. The state of the art. In J.M. Tanner (Ed.). *Auxiology '88: Perspectives in the science of growth and development* (pp. 1-29). London: Smith-Gordon.

Hirose, K. (2003). Comparing L1 and L2 organizational patterns in the argumentative writing of Japanese EFL students. *Journal of second language writing, 12*, 181-209.

Janssen, T., Braaksma, M., Rijlaarsdam, G., & Van den Bergh, H. (2005). *Flexibility in reading literary texts; differences between weak and strong adolescent readers.* Paper presented at the 11th European Conference for Learning and Instruction. Nicosia, Cyprus, August 22-27, 2005.

Jöreskog, K.G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36* (2), 109-133.

Kieft, M., Rijlaarsdam, G., & Van den Bergh, H. (2006). Writing as a learning tool. Testing the role of students' writing strategies. *European Journal of Psychology of Education*, *21* (1), 17-34.

Kieft, M., Rijlaarsdam, G., & Van den Bergh, H. (2008). An aptitude-treatment interaction approach to writing-to-learn. *Learning and Instruction*, *18*, 379-390.

Kuhlemeier, H., & Van den Bergh, H. (1998). Relationships between language skills and task effects. *Perceptual and Motor Skills*, *86*, 443-463.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: size, strength, and computer adaptiveness. *Language Learning*, *54* (3), 399-436.

Lavelle, E., Smith, J., & O'Ryan, L. (2002). The writing approaches of secondary students. *British Journal of Educational Psychology*, *72*, 399-418.

Leijten, M., & Van Waes, L. (2006). Inputlog: New perspectives on the logging of on-line writing. In G. Rijlaarsdam (Series Ed.), K.P.H. Sullivan, & E. Lindgren (Vol. Eds.). *Studies in Writing. Vol. 18. Computer key-stroke logging and writing: methods and applications* (pp. 73-94). Oxford: Elsevier

Lindgren, E., Spelman Miller, K., & Sullivan, K. (2008). Development of fluency and revision in L1 and L2 writing in Swedish high school years eight and nine. *ITL Applied Linguistics*, *156*, 133-151.

McCutchen, D. (1996). A capacity theory of writing: working memory in composition. *Educational Psychology Review*, *8* (3), 299-325.

Olive, T., Kellogg, R., & Piolat, A. (2008). Verbal, visual, and spatial working memory demands during text composition. *Applied Psycholinguistics*, *29*, 669-687.

Purves, A.C. (Ed.) (1992). *The IEA study of written composition II: Education and performance in fourteen countries*. Oxford: Pergamon Press.

Purves, A.C., Gorman, T.P., & Takala, S. (1988). The development of the scoring scheme and scales. In T.P. Gorman, A.C. Purves, & R.E. Degenhart (Eds.). *The IEA study for written composition I : the international writing tasks and scoring scales* (pp. 41-58). Oxford: Pergamon Press.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective. *Language Learning*, *52* (3), 513-536.

Raymond, J.C. (1982). What we don't know about the evaluation of writing. *College Composition and Communication*, *33* (4), 399-403.

Rijlaarsdam, G., & Van den Bergh, H. (1996). The dynamics of composing. An agenda for research into an interactive compensatory model of writing. Many questions, some answers. In C.M. Levy, & S.Ransdell (Eds.). *The science of writing. Theories, methods, individual differences, and applications* (pp. 107-125). Mahwah, N.J.: Lawrence Erlbaum Associates.

Rijlaarsdam, G., Van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Van Steendam, E., & Raedts, M. (2011). Writing. In K.R. Harris, S. Graham, & T. Urdan (Eds.). APA *Educational Psychology Handbook. Vol. 3* (pp. 189-227). Washington: American Psychological Association.

Roca de Larios, J., Manchón, R., & Murphy, L.(1996). Strategic knowledge in L1 and L2 writing: a cross-sectional study. Barcelona: Presentation at EARLI SIG Writing Conference, 23rd-25th October.

Roca de Larios, J., Marín, J., & Murphy, L. (2001). A temporal analysis of formulation processes in L1 and L2 writing. *Language Learning*, *51*, 497-538.

Roca de Larios, J., Manchón, R., Murphy, L., & Marín, J. (2008). The foreign language writer's strategic behaviour in the allocation of time to writing processes. *Journal of Second Language Writing*, *17* (1), 30-47.

Russo, J.E., Johnson, E.J., & Stephens, D.L. (1989). The validity of verbal protocols. *Memory & Cognition*, *17* (6), 759-769.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning*, *46* (1), 137-174.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation modeling. *Language Testing*, *22* (1), 1-30.

Schoonen, R., Van Gelderen, A., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2003). First language and second language writing: the role of linguistic knowledge, speed of processing and metacognitive knowledge. *Language Learning*, *53* (1), 165-202.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: the ESL research and its implications. *TESOL Quarterly*, *27* (4), 657 - 677.

Stevenson, M., Schoonen, R., & De Glopper, K. (2006). Revising in two languages: a multi-dimensional comparison of online writing revisions in L1 and FL. *Journal of Second Language Writing*, *15*, 201-233.

Strömqvist, S., Holmqvist, K., Johansson, V., Karlsson, H., & Wengelin, Å. (2006). What key-logging can reveal about writing. In K. Sullivan, & E. Lindgren (Eds.). *Studies in Writing. Vol. 18. Computer key-stroke logging and writing: methods and applications* (pp. 45-72). Boston: Kluwer Academic Publishers.

Thorson, H. (2000). Using the computer to compare foreign and native language writing processes: a statistical and case study approach. *The Modern Language Journal*, *84* (2), 155-170.

Torrance, M., Thomas, G.V., & Robinson, E.J. (1994). The writing strategies of graduate research students in the social sciences. *Higher Education*, *27*, 379-392.

Torrance, M., Thomas, G.V., & Robinson, E.J. (1999). Individual differences in the writing behaviour of undergraduate students. *British Journal of Educational Psychology*, *69*, 189-199.

Torrance, M., Thomas, G.V., & Robinson, E.J. (2000). Individual differences in undergraduate essay-writing strategies. A longitudinal study. *Higher Education*, *39*, 181-200.

Uzawa, K. (1996). Second language learners' processes of L1 writing, L2 writing and translation from L1 into L2. *Journal of Second Language Writing*, *5* (3), 271-294.

Van den Bergh, H. (1988a). *Examens geëxamineerd. [Exams examined.]* 's-Gravenhage, the Netherlands: SVO.

Van den Bergh, H. (1988b). Schrijven en schrijven is twee: Een onderzoek naar de samenhang tussen prestaties op verschillende schrijftaken. [Writing and writing makes two: An investigation into the relation between performances on different writing tasks.] *Tijdschrift voor Onderwijsresearch*, *13* (6), 311-324.

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

Van den Bergh, H., & Rijlaarsdam, G. (1996). The dynamics of composing. Modelling writing process data. In C.M. Levy & S.Ransdell (Eds.). *The science of writing. Theories, methods, individual differences, and applications* (pp. 207-232). Mahwah, N.J.: Lawrence Erlbaum Associates.

Van den Bergh, H., & Rijlaarsdam, G. (1999). The dynamics of idea generation during writing: an online study. In G. Rijlaarsdam & E. Espéret (Series Eds.), M. Torrance & D. Galbraith (Vol. Eds.). *Studies in Writing. Vol. 4. Knowing what to write: conceptual processes in text production* (pp. 99-120). Amsterdam: Amsterdam University Press.

Van den Bergh, H., & Rijlaarsdam, G. (2001). Changes in cognitive activities during the writing process and relationships with text quality. *Educational Psychology*, *21* (4), 373-385.

Van den Bergh, H., & Klein Gunnewiek, L. (2009). CEF of cijfers bij de beoordeling van schrijfvaardigheid. [CEF or marks for judging writing ability.] *Levende Talen Tijdschrift*, *10*, 3-11.

Van den Bergh, H., Rijlaarsdam, G., Janssen, T., Braaksma, M., Van Weijen, D., & Tillema, M. (2009). Process execution of writing and reading. Considering text quality, learner and task characteristics. In M.C. Shelley II, L.D. Yore & B. Hand (Eds.). *Quality research in literacy and science education* (pp. 399-425). Dordrecht, The Netherlands: Springer.

Van der Hoeven, J. (1997). *Children's composing: A study into the relationship between writing processes, cognitive and linguistic skill and text quality.* Amsterdam: Rodopi.

Van der Stel, M., & Veenman, M.V.J. (2008). Relation between intellectual ability and metacognitive skillfulness as predictors of learning performance of young students performing tasks in different domains. *Learning and Individual Differences*, *18*, 128-134.

Van Weijen, D. (2009). *Writing processes, text quality, and task effects: Empirical studies in first and second language writing.* Utrecht: LOT Dissertation Series.

Van Weijen, D., Van den Bergh, B., Rijlaarsdam, G., & Sanders, T. (2008). Differences in process and process-product relations in L2 writing. *ITL Applied Linguistics*, *156*, 203-226.

Van Weijen, D., Van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2009). Composing episodes. Exploring the hierarchical structure of L1 and L2 writing. Manuscript submitted for publication.

Veenman, M.V.J., Prins, F.J., & Verheij, J. (2003). Learning styles. Self-reports versus thinking-aloud measures. *British Journal of Educational Psychology*, *73*, 357-372.

Veenman, M.V.J., & Spaans, M.A. (2005). Relation between intellectual and metacognitive skills: Age and task differences. *Learning and Individual Differences*, *15*, 159-176.

Weigle, S.C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

Wesdorp, H. (1974). Het meten van de 'stelvaardigheid' [Measuring 'writing ability']. *Pedagogische Studiën*, *51*, 499-521.

White, E.M. (1984). Holisticism. *College Composition and Communication*, *35* (4), 400-409.

White, E.M. (1985). *Teaching and assessing writing*. San Francisco, CA: Jossey-Bass.

Zohar, A., & David, A.B. (2009). Paving a clear path in a thick forest. A conceptual analysis of a metacognitive component. *Metacognition and Learning*, *4*, 177-195.

APPENDIX A

(chapter 2, p. 15; chapter 3, p. 40; chapter 4, p. 64; chapter 5, p. 86)

Example of an assignment.

Surveillance cameras in inner city areas

NACP, the National Action Committee for Pupils, is organising a national essay contest, especially for pupils of your age. You're also taking part. You absolutely want to win. The winning essay will be printed in PAUZE, a monthly magazine that is read by pupils your age from all over the Netherlands.

The subject of the essay has already been decided and was described in PAUZE as follows:
Due to the increase in crimes and meaningless acts of violence, more and more cities are choosing to place surveillance cameras in inner city areas. Not everyone is pleased about this. Some feel safe knowing that someone is 'watching over' them, while others consider it an invasion of their privacy. NACP is going to pay attention to this subject in a special edition of PAUZE. We want to hear from pupils what they think. Decide what you think and send us your response!

Assignment:
Write an essay in which you give your opinion on the question:
"Do surveillance cameras in inner city areas increase public security?"

The essay has to meet the following requirements, set by the Jury:

1. Your essay must be (about) half a page in length.
2. You must do your best to convince your readers, fellow pupils, of your opinion.
3. You must give arguments to support your opinion.
4. Your essay must be structured in a good and logical way.
5. Your essay must look well-cared-for (think of language use and spelling).
6. In your essay you must use at least two extracts from the 'References' (see next page). You must include these extracts in your essay in a meaningful way.

You have 30 minutes to complete this assignment.

Good luck!

References

Surveillance cameras help prevent crime, but they also increase the chance of tracking down the culprits. […] The cameras are placed in such a way that perpetrators of crimes within the inner city area are almost always registered. […] Incidentally, despite the presence of surveillance cameras, the security and well-being of the general public remains everyone's concern. That's why we still say: "if you spot trouble, warn the police!"
Source: Maastricht County Council, www.maastricht.nl, 2004.

The evaluation report of the project in Ede already mentioned that surveillance cameras don't just take away 'feelings of unease and insecurity.' On the contrary, before the cameras were installed, 65% of people visiting the Museumplein never felt unsafe, whereas five months after installation the percentage had dropped to 57%. Scottish research has also shown that after a short decrease the 'feelings of insecurity' rise again.
Source: Erik Timmerman, Leeuwarder Courant, July 7th 2000.

Great Britain has become THE surveillance capital of Europe, without anyone noticing, says Barry Hugill, spokesman for the English civil rights group Liberty. Remarkably, the call for privacy is gradually being overshadowed by experts' warning against 'a false sense of security.' Ian Brown, researcher for Information Policy Research […]: "It is an illusion to think that cameras will provide security." […] One study showed that in areas with intensive camera surveillance crime rates dropped by 3 or 4 percent, whereas better street lighting can help reduce the number of incidents by up to 20 percent. The general public is usually less vigilant, as the number of cameras in the area increases.
Adapted from: Steven de Jong, www.politiek-digitaal.nl, August 30th 2004.

The crime rate on the Wallen and in the vicinity of the Nieuwendijk in the centre of Amsterdam has decreased since the implementation of camera surveillance in March. […] In the Nieuwendijkkwartier especially, satisfaction prevails all round. […] In the vicinity of the Wallen, people are generally positive, although attention is drawn to the unwanted relocation of problems to other areas and changes in the group of troublemaking drug dealers and junkies. To combat these effects, the council will install three extra cameras.
Adapted from: Centrum voor Criminaliteitspreventie en Veiligheid, www.ccv.nu, November 30th 2004.

Security cameras in and around the Noorderstation are functioning with difficulty. It is even questionable whether the cameras, installed last year, are actually working. […] Not all the cameras are permanently on-line […] police spokesman Ed Kraszewski reporterd last week. They are chiefly there as a preventative measure. He believes that the presence of cameras will deter thieves and violent criminals from committing criminal acts, even if the cameras are not on-line.
Source: Cees Vellekoop, Dagblad van het Noorden, August 18th 2003.

Permanent camera surveillance in Sneek's inner city is pointless. […] The city council based her decision, amongst other things, on the experiences of other city councils with security cameras. These experiences taught them that not all cities have had positive results and that the costs, especially personnel costs, have been substantial.
Adapted from: Friesch Dagblad, April 6th 2004.

APPENDIX B (chapter 2, p. 16)

Definitions of rating criteria, as provided to the raters.

Global quality
General impression of the quality of the essay's content, persuasiveness, reasoning, argumentation, goal orientation, reader orientation, language use and appearance.

Structure
- Quality of reasoning and argumentation
- Quality of the manner in which the essay's build-up and appearance supports reasoning and argumentation.

Language
- Spelling
- Style, i.e.
    *accuracy of formulated language ("to-the-point-ness")
    *suitability of vocabulary and sentence complexity for intended readers and medium
    *originality of language use
- Tone (suitability for intended readers and medium)

APPENDIX C (chapter 2, p. 17)

Example of a benchmark essay (rating criterion structure; L2 benchmark) and the accompanying description of what is ‘average’ about it.

> Living alone yes or no?
>
> The descision of living, together, get married or stay single should everyone descide for themselves. Because everyone has other needs. Some people like to live alone, and other people need someone around.
> I think marridge isn't a must, a person can live alone very well, without feeling miserable, but of coure not everyone.
> Marridge should not be rushed, you should think it over very well, and wait for the perfect partner for you! And before you get married you should know eachother very well.
> I want to get married in the future but first, the right man has to come along. I also think it's better if you first live together and when it goes well, you can get married. So you will find out how living with each other in one house is.
> I just think it would be very nice to live with the person you love, and see each other every day, especially waking up together each morning would be nice.
>
> Quote: We expect single people to be somewhat lonely and unhappy, and we expect people who are married or who live together to be absolutely happy.
>
> I think this is not true.
> People who choose to live alone can be very happy, it depend on the person. And people who got married and have picked the right person can be happy too.
> I think marridge is fine, but it should not be rushed, and if you live alone you can still get married as well.

Explanation: what is average about this benchmark essay's structure?
Its strong points in terms of structure:
* The main statement of this essay is clearly stated in the first paragraph.
* Each paragraph pertains to one single idea/argument.
* For every new or separate idea, there is a new paragraph.
* Within paragraphs, there is a consistent line of reasoning.
Its weak points in terms of structure:
* Not every paragraph supports the main statement.
* Some paragraphs interrupt the line of reasoning at text level.
* Transitions are not clearly marked. The quote, for example, is not introduced. Because the quote is unrelated to the previous paragraph, the reader gets side-tracked.

APPENDIX D (chapter 2, p. 18; chapter 3, p. 45)

Procedural manual

For this manual, we will work with the numbers of essays, participants and raters as used in the present study. The procedure can be adjusted to different amounts. To be able to make general claims about an individual's writing proficiency in each of the languages under investigation, multiple tasks should be used per language.

A. Obtaining essays

1) Create eight writing assignments, which differ as slightly as possible. For instance, the essays differ only in terms of topic: topics A, B, C, D, E, F, G, H.
2) Each of the eight assignments should be available in both L1 and L2.
3) Balance the topics across participants and experimental conditions (i.e. language in which the text is written) as follows:

Table D1. Balancing topics (A, B, …, H) across participants and experimental conditions

Participant	Participant writes in L1	Participant writes in L2
1	A, B, C, D	E, F, G, H
2	B, C, D, E	F, G, H, A
3	C, D, E, F	G, H, A, B
4	D, E, F, G	H, A, B, C
5	E, F, G, H	A, B, C, D
6	F, G, H, A	B, C, D, E
etc.		

4) The L1 and L2 writing sessions (= experimental conditions) are administered on different days. Balance the order in which experimental conditions are presented as follows:

Table D2. The order of conditions per participant

Participant	Day 1	Day 2
1	L1 (topics A, B, C, D)	L2 (topics E, F, G, H)
2	L2 (topics F, G, H, A)	L1 (topics B, C, D, E)
3	L1 (topics C, D, E, F)	L2 (topics G, H, A, B)
4	L2 (topics H, A, B, C)	L1 (topics D, E, F, G)
5	L1 (topics E, F, G, H)	L2 (topics A, B, C, D)
6	L2 (topics B, C, D, E)	L1 (topics F, G, H, A)
etc.		

B. Creating subsamples

5) Mix the essays and create eight subsamples, which consist of twenty randomly selected essays each. Generally, each subsample will now contain both L1 and L2 essays on various topics, written by various participants.
6) Number the essays in such a way that the raters cannot trace them back to specific authors. For example: subsample 1 contains essays 1 through 20, subsample 2 contains essays 21 through 40, and so forth. The researcher will keep a file where the new numbers are matched to the original participant and task numbers.

C. Selecting benchmarks essays

7) Select six benchmark essays:

* an L1 benchmark for global quality
* an L1 benchmark for structural quality
* an L1 benchmark for language quality
* an L2 benchmark for global quality
* an L2 benchmark for structural quality
* an L2 benchmark for language quality

(Other criteria of text quality could be used, but there must be multiple criteria.)
8) The benchmark essays represent (approximate) average quality for each of the specific rating criteria. Which essays qualify as average, can be determined on the basis of a pre-rating session.

D. Ratings

9) Select raters who are highly and equally proficient in both L1 and L2, and aware of text conventions in each of the two languages. Ideally, they should be bilinguals.
10) These raters cannot have been involved in the pre-rating session to select benchmark essays.
11) The ratings for global quality are performed first. To minimize the possibility of occurrence of halo effects, at least a week should pass before the 'structure' ratings take place, and another week after that before the 'language' ratings are performed.
12) Each rater rates three subsamples relative to the L1 benchmark for global quality, and three subsamples relative to the L2 benchmark for global quality:
13) Balance the order in which benchmark languages occur: 4 raters perform ratings with L2 benchmarks first, and ratings with L1 benchmarks second, and the 4 remaining raters perform ratings with L1 benchmarks first, and ratings with L2 benchmarks second.
14) Provide the raters with a definition of the rating criterion, along with a description of what is average about the benchmark essay (Appendices B and C).

Table D3. Distribution of subsamples across raters[11]
L1 = rating with L1 benchmark; L2 = rating with L2 benchmark

	Subsample							
	1	2	3	4	5	6	7	8
Rater 1	L1	L1	L1	L2	L2	L2		
Rater 2		L1	L1	L1	L2	L2	L2	
Rater 3			L1	L1	L1	L2	L2	L2
Rater 4	L2			L1	L1	L1	L2	L2
Rater 5	L2	L2			L1	L1	L1	L2
Rater 6	L2	L2	L2			L1	L1	L1
Rater 7	L1	L2	L2	L2			L1	L1
Rater 8	L1	L1	L2	L2	L2			L1

15) The benchmark essays are awarded the randomly set score of 100. Instruct the raters to award a score to each essay which expresses how much better or worse they think it is than the benchmark essay in terms of global quality. If an essay was awarded a score of 200, for example, this meant that the rater thought its global quality was twice as good as the benchmark essay. If an essay received a score of 50, it meant that the rater thought it was half as good as the benchmark essay.
16) Steps 12, 13, 14 and 15 are repeated during the 'structure' and 'language' ratings. The only differences will be the rating criterion and (therefore) the benchmark essays.

[11] As the reader can see in table 1 of chapter 2, the advised allocation of subsamples to raters and benchmarks, as presented in table D3, was not completely followed in the reported study. This is due to a small logistic mishap, resulting in the need to shuffle the allocation of subsamples to raters (as compared to the advised allocation in table D1). Nevertheless, the set-up presented in table 1 still suits the requirements of the study, namely that:
A) Each subsample was rated three times relative to an L1 benchmark and three times relative to an L2 benchmark.
B) Each rater rated three subsamples relative to an L1 benchmark and three subsamples relative to an L2 benchmark.
C) Each rater therefore rates each essay only once (for each criterion). That is, he or she never rates the same essay relative to both an L1 and an L2 benchmark.

APPENDIX E (chapter 3, p. 48)

Effects of episode and language on the percentage with which cognitive activities occur. RA = Reading Assignment; PP = Process Planning; CP = Content Planning; F = Formulating; ROT = Reading Own Text; EOT = Evaluating Own Text; RV = Revising.

	Main effects		**Interaction effect**
	Episode	**Language**	**Episode*language**
RA	F (4, 681.6) = 69.6; p < .001	F (1, 19.1) = .002; p = .97	F (4, 681.6) = 1.4; p = .23
PP	F (4, 678.4) = 24.2; p < .001	F (1, 17.0) = .28; p = .60	F (4, 678.5) = 3.3; p < .05
CP	F (4, 672.4) = 15.3; p < .001	F (1, 19.3) = 5.0; p < .05	F (4, 672.4) = 0.7; p = .59
F	F (4, 679.3) = 27.4; p < .001	F (1, 18.2) = 3.9; p = .06	F (4, 679.3) = 2.7; p < .05
ROT	F (4, 680.7) = .00; p = 1.00	F (1, 30.3) = .12; p = .73	F (4, 680.7) = .00; p = 1.00
EOT	F (4, 692.2) = 15.6; p < .001	F (1, 19.2) = 0.4; p = .52	F (4, 692.2) = 0.6; p = .68
RV	F (4, 679.4) = .00; p = 1.00	F (1, 29.2) = 3.3; p = .08	F (4, 679.4) = .00; p = 1.00

APPENDIX F (chapter 3, p. 50)

Relations between cognitive activities and text quality per episode: parameter estimates. Significant estimates ($p < .05$) are marked with *.
The estimates in the L1→L2 columns indicate whether the size of the relation between text quality and the cognitive activity in this episode changes in L2 relative to L1. For example, Reading the Assignment is positively related to L1 text quality if it occurs during episode 2: *β2* = 1.08 (significant). Or: the more Reading the Assignment occurs (in percentages) in episode 2 during L1 writing tasks, the higher the text quality score. The estimate in 'L1→ L2' shows that the effect of Reading the Assignment during episode 2, which was positive during L1 writing, becomes smaller if writing occurs in L2: *β7* = -2.06 (significant). The actual L2 relation between Reading the Assignment in episode 2 is obtained as follows: *β2*- *β7* = 1.08 – 2.06 = -0.98 (the significance level of the L2 estimates was tested). This relation is negative: the more Reading the Assignment occurs in episode 2 during L2 writing tasks, the lower the text quality score.

	Reading the Assignment					
	L1		**L1→ L2**		**L2**	
Intercept	*β0*	117.28*				
Episode 1	*β1*	0.64	*β6*	-1.31*	*β1*- *β6*	-0.67*
Episode 2	*β2*	1.08*	*β7*	-2.06*	*β2*- *β7*	-0.98*
Episode 3	*β3*	-0.03	*β8*	-0.55	*β3*- *β8*	-0.58
Episode 4	*β4*	-0.23	*β9*	-0.53	*β4*- *β9*	-0.76
Episode 5	*β5*	0.00	*β10*	0.19	*β5*- *β10*	0.19

	Process Planning					
	L1		**L1→ L2**		**L2**	
Intercept	*β0*	118.05				
Episode 1	*β1*	0.00	*β6*	-2.09*	*β1*- *β6*	-2.09*
Episode 2	*β2*	0.59	*β7*	-0.13	*β2*- *β7*	0.46
Episode 3	*β3*	2.93*	*β8*	-1.75	*β3*- *β8*	1.18
Episode 4	*β4*	-2.07*	*β9*	0.11	*β4*- *β9*	-1.96*
Episode 5	*β5*	2.83*	*β10*	-5.30*	*β5*- *β10*	-2.47*

	Content Planning					
	L1		**L1→ L2**		**L2**	
Intercept	*β0*	111.64				
Episode 1	*β1*	2.66*	*β6*	-3.04*	*β1*- *β6*	-0.38
Episode 2	*β2*	-1.94	*β7*	1.73	*β2*- *β7*	-0.21
Episode 3	*β3*	4.02*	*β8*	-3.56*	*β3*- *β8*	0.46
Episode 4	*β4*	-0.21	*β9*	-2.35	*β4*- *β9*	-2.56*
Episode 5	*β5*	-0.72	*β10*	1.15	*β5*- *β10*	0.43

	Formulating					
	L1		**L1→ L2**		**L2**	
Intercept	*β0*	110.27				
Episode 1	*β1*	0.18	*β6*	-0.15	*β1- β6*	0.03
Episode 2	*β2*	0.14	*β7*	-0.23	*β2- β7*	-0.09
Episode 3	*β3*	0.40	*β8*	-0.51	*β3- β8*	-0.11
Episode 4	*β4*	0.19	*β9*	0.32	*β4- β9*	0.51
Episode 5	*β5*	-0.34	*β10*	-0.46	*β5- β10*	-0.80*
	Reading Own Text					
	L1		**L1→ L2**		**L2**	
Intercept	*β0*	112.78				
Episode 1	*β1*	0.84	*β6*	-1.97	*β1- β6*	-1.13
Episode 2	*β2*	-0.56	*β7*	-0.12	*β2- β7*	-0.68
Episode 3	*β3*	-0.40	*β8*	-0.48	*β3- β8*	-0.88
Episode 4	*β4*	1.77*	*β9*	-2.37*	*β4- β9*	-0.60
Episode 5	*β5*	1.49*	*β10*	-2.20*	*β5- β10*	-0.71
	Evaluating Own Text					
	L1		**L1→ L2**		**L2**	
Intercept	*β0*	116.67				
Episode 1	*β1*	-6.22	*β6*	3.11	*β1- β6*	-3.11
Episode 2	*β2*	4.47*	*β7*	-11.32*	*β2- β7*	-6.85*
Episode 3	*β3*	-1.20	*β8*	5.79	*β3- β8*	4.59*
Episode 4	*β4*	3.53*	*β9*	-6.63*	*β4- β9*	-3.10*
Episode 5	*β5*	2.33*	*β10*	-7.62*	*β5- β10*	-5.29*
	Revising					
	L1		**L1→ L2**		**L2**	
Intercept	*β0*	104.36				
Episode 1	*β1*	0.42	*β6*	-0.67	*β1- β6*	-0.25
Episode 2	*β2*	0.51	*β7*	-0.78	*β2- β7*	-0.27
Episode 3	*β3*	1.26*	*β8*	-1.55*	*β3- β8*	-0.29
Episode 4	*β4*	0.73	*β9*	-1.50*	*β4- β9*	-0.77
Episode 5	*β5*	0.72*	*β10*	-0.71	*β5- β10*	0.01

APPENDIX G (chapter 4, p. 68)

L1 language proficiency test:

	Sentence	**Solution**
1	Ik kwam Denise gisteren nog ... in de supermarkt.	tegen
2	Ik heb een nieuwe fiets. Ik ... er ontzettend blij mee.	ben
3	Hoe ... ken je Bob al? - Ik ken hem al negen jaar.	lang
4	... was je gisteren niet op school? - Ik was gisteren niet op school, omdat ik ziek was.	waarom
5	Het is vandaag precies 400 jaar … dat de telescoop werd uitgevonden.	geleden
6	Ik heb je gedrag tot nu toe getolereerd, maar nu ben je toch echt te ... gegaan.	ver
7	Mevrouw, u vergeet ... tas!	uw
8	Mijn grootouders hebben hun hele ... hard gewerkt en genieten nu van hun pensioen.	leven
9	Kom je aan tafel? Het eten staat al ...	klaar
10	Toen we jonger waren, maakten mijn zusje en ik vaak ... Maar tegenwoordig kunnen we erg goed met elkaar opschieten.	ruzie
11	Ik ben het nog steeds niet met hun beslissing eens, maar ik heb besloten me erbij neer te …	leggen
12	Mijn moeder is de liefste moeder van de hele ...	wereld
13	Ik wil graag op reis naar Bolivia, ... ik heb er het geld niet voor.	maar
14	Er zijn steeds meer vrouwelijke vrachtwagenchauffeurs. Maar bij de meeste trucks zit er nog steeds een man ... het stuur.	achter
15	Mark is op zijn kamer. Wil jij deze boeken bij ... brengen?	hem
16	John beloofde om voor ... en eeuwig bij haar te blijven.	altijd
17	Tegenwoordig heeft iedereen een wasmachine, maar dat was ... wel anders.	vroeger
18	Ik maak een notitie in mijn agenda, ... ik onze afspraak niet kan vergeten.	zodat
19	Het ozongat boven het Zuidpoolgebied blijft groeien. Dit jaar is het ozongat ... dan ooit werd gemeten.	groter

20	Mijn koffer is bijna gepakt. Ik moet alleen mijn toiletspullen … inpakken.	nog
21	Ik had geen idee wat ik mijn vriendin voor haar verjaardag moest geven, maar gisteravond kreeg ik ... een goed idee.	opeens, ineens, toch, plotseling, gelukkig
22	Het is verplicht om in de fabriek veiligheidsschoenen te ...	dragen
23	De ochtendkrant meldde dat de gijzelnemers morgen een aantal gijzelaars ... zullen laten.	vrij
24	Bart denkt nooit aan anderen. Hij is alleen maar met ... bezig.	zichzelf
25	Ik weet niet precies hoe oud Julia is, maar ik denk dat ze ... 40 is.	ongeveer/circa
26	Ik vind mijn huidige woonplaats niet leuk meer. Ik wil graag ... anders wonen.	ergens
27	Je moet echt beter je best doen op school. ... zul je dit jaar niet overgaan naar de volgende klas.	anders
28	Er waren maar drie bekenden van mij op het feest, verder kende ik er ...	niemand
29	Een keelontsteking gaat vrijwel altijd zonder medicijnen weer over. Heel ... is het echter nodig om medicijnen te nemen.	soms
30	De wedstrijd Ajax-Feyenoord eindigde in ... spel. Het werd 2-2.	gelijk
31	Ben je nu pas aan je huiswerk begonnen? Dat had je veel ... moeten doen!	eerder
32	Zij is niet van Nederlandse afkomst. - Waar komt ze dan ...?	vandaan
33	De rechtbank van Leeuwarden … later vandaag uitspraak in de Marssumse moordzaak.	doet
34	Omdat we de anderen niet wakker wilden maken, probeerden we zo ... mogelijk te praten.	zacht(jes)
35	Maria is op dit … niet aanwezig. Kun je over een uurtje terugbellen?	Moment/ogenblik
36	Mijn broer is net vader ... Zijn kindje is vorige week geboren.	geworden
37	Ik weet nog niet of ik morgen naar het parkfestival ga. Het … ervan af of het wel of niet gaat regenen.	hangt
38	Ik wilde vorige maand graag naar het concert van mijn	mocht

	favoriete band, maar het ... niet van mijn ouders.	
39	De ambulance was na het ongeval snel ... plaatse.	ter
40	Er zijn veel landen in Europa die ik nog nooit heb bezocht. Ik ben, ..., nog nooit in Zwitserland geweest.	bijvoorbeeld
41	Het ... niet vaak, maar soms wordt er een ijsberendrieling geboren.	gebeurt
42	Zijn dit de jassen van Sylvia en Martin? - Nee, ... jassen hangen aan de kapstok.	hun
43	Eerlijk ... vind ik jouw nieuwe schoenen niet zo mooi.	gezegd
44	Het bedrijf bood Jan een ontslagvergoeding van twee maanden aan, maar daar … hij geen genoegen mee.	nam
45	De groene mamba is de … dodelijke slang ter wereld. Na een beet kun je binnen een half uur dood zijn.	meest
46	Ik had vorige week vier proefwerken. Ze gingen allemaal goed, ... het wiskundeproefwerk. Ik vrees dan ook dat ik voor wiskunde een onvoldoende zal halen.	behalve
47	Hoewel mijn broer de schuldige was, ... mijn vader mij de schuld.	gaf
48	Het meer is hier niet erg ... Het water komt hooguit tot aan mijn knieën.	diep
49	Het feest was een groot succes, met ... aan alle vrijwilligers die hebben geholpen.	dank
50	Ik weet dat je graag wilt dat ik met je mee ga. Maar ik heb er gewoon geen zin in en … basta!	daarmee
51	Je mag meedoen aan deze cursus, maar het ... niet per se.	hoeft
52	Je mag hier niet sneller rijden dan 50 ... per uur.	kilometer
53	Aanvankelijk vond hij het een goed plan, maar hij heeft zich ... Hij wil bij nader inzien toch niet meedoen.	bedacht
54	Zij is niet in Utrecht geboren, maar woont hier al ... 1960.	sinds
55	Utrecht ligt ten ... van Tilburg, maar ten zuiden van Amsterdam.	noorden
56	Hoewel er vandaag in heel Nederland kans is op regen, zal er … in het noorden en oosten van het land veel regen vallen. In het westen en zuiden zal het droger blijven.	vooral
57	Je stelt een goede vraag. Helaas kan ik je geen ...	antwoord

	geven.	
58	Ik heet Marieke, maar mijn moeder … me altijd Riekje.	noemt/noemde
59	Op welke politieke ... ga jij stemmen bij de volgende verkiezingen?	partij
60	90% van de leerlingen vindt het belangrijk dat een docent goed ... kan houden. Ze hebben er een hekel aan als het een chaos is in de klas.	orde
61	De kozijnen zijn gaan rotten, ... ze vijftien jaar lang niet opnieuw geverfd zijn.	doordat/nadat
62	... hij een druk leven leidt, maakt hij altijd tijd voor me vrij.	hoewel

L2 language proficiency test:

	Sentence	Solution
1	It is not ... [moeilijk] to see why he did not recognize you.	hard/difficult
2	I went to the school canteen ... [nadat] I had been sent out of the classroom.	after
3	Do girls ... [voetballen] at your school?	play football / soccer
4	... [Als] I had enough money I would buy a walkman.	If
5	There were some riots ... [buiten] the South-African Embassy.	outside
6	My granddad died during the war ... [op] the age of thirty-eight.	at
7	I hope you'll write ... [me terug] soon.	(me) back / back to me
8	... [Eerst] I shall tell you about my family and then about my friends.	First
9	As ... [gewoonlijk], I had forgotten the frontdoor-key when I came home late.	usual
10	I left primary school ... [op] the age of eleven.	at
11	Barbara drove as ... [hard] as she could.	fast
12	I ... [vind] that we should go to Percy's funeral tomorrow.	think
13	Well, I must ... [beëindigen] this letter now.	end
14	... [Vraag het] your teacher.	Ask
15	I spent all night ... [naar de televisie te kijken].	watching TV/watching television
16	It ... [maakt niet uit] which date you choose.	doesn’t matter
17	You can ... [lenen] my tent if you want to.	borrow
18	In our shop we have quite a lot of newspapers, The Observer, ... bijvoorbeeld], and The Times.	for example/for instance
19	Your letter was shorter than ... [gebruikelijk].	usual
20	I think I'm going to apply ... [naar] a holiday job.	for
21	How did your little sister manage to get ... [uit] bed?	out of
22	You can sleep in my tent if you want to. There is ... [plaats] for three.	room
23	I read ... [de meeste] of Agatha Christie's detective novels.	most

24	At school I am not allowed to sit ... [naast] my friend.	next to/beside
25	The pupils all sit ... [aan] tables when doing their exams.	at
26	There is a ... [verwarmd] swimming-pool at the camping site.	heated
27	We've been waiting ... [al] more than an hour.	for
28	I have ... [mijn hoofdpijn kwijt] at last.	lost my headache/ got rid of my headache
29	... [Als] child I often stayed with my grandparents.	As a
30	I didn't realize you had been waiting for an answer ... [van] me all the time.	from
31	I'll finish grammar school first and ... [dan] go to university.	then
32	Our class visited a kennel where guide dogs are trained to lead the ... [blinden].	blind
33	We were already late for the cinema when our car ... [kapot ging].	broke down
34	Read ... [door], please, my letter will become more interesting.	on
35	It was almost half an hour ... [eer] the fire brigade arrived.	before
36	We were all late for school, because we had ... [ons verslapen].	overslept
37	I usually watch TV ... [op] my own room.	in
38	I have written this essay all ... [alleen].	by myself
39	I usually ... [maak] my homework right after school.	do
40	... [Zou je niet willen] to come and stay with us next weekend?	Wouldn't you like
41	Is there ... [iets] I can do to help you?	anything
42	The first morning I felt sick, so I didn't want to ... [me niet aankleden], but stay in bed all day.	get dressed
43	I hope there will be a ... [rechtstreeks] flight to Birmingham.	direct
44	My brother will have to stay ... [in het] hospital for ten days.	in
45	We met him ... [in] Baker Street.	in
46	... [Wat] is further from your home: the bus stop or the railway station?	Which

47	I have no money ... [bij] me.	on / with
48	Rewrite the first ... [alinea].	paragraph
49	I was looking for a table ... [waarop] I could put my books.	on which
50	Our school is ... [bij] the railway station.	near
51	We had to write an essay on ... [rassen] discrimination.	race / racial
52	My ... [oudste] sister is getting married next Friday.	eldest
53	My uncle is ... [leraar].	a teacher
54	I had to tell the psychologist what things I took an interest ... [voor].	in
55	My father is ... [voorzitter] of the parents' council.	chairman / head / president
56	I'm looking ... [er naar uit] seeing you soon.	forward to
57	I was woken by the light of a torch that ... [scheen] in my face. When I opened my eyes I saw a man pointing a gun at me.	shone
58	I speak ... [Frans noch Spaans], I only speak English.	neither French nor Spanish
59	You don't know ... [hoe het is] to be alone.	what it's like
60	As there was no bus, we had to go all the way ... [te voet].	on foot
61	I realized too late that I was sitting on a ... [pas] painted bench.	freshly / newly / recently
62	I ... [wandel niet graag] through the park on my own.	don't like walking
63	I have to practise the piano every day ... [of] I like it or not.	whether
64	We had thirty questions ... [in totaal] to answer.	in all

APPENDIX H (chapter 4, p. 70)

Regression model used for explaining the occurrence of each of the seven cognitive activities:

$$A_{ij} = \beta_{0ij} * epi^{0}_{ij} + \beta_{1i} * epi^{1}_{ij} + \beta_{2} * epi^{2}_{ij} + \beta_{3} * epi^{3}_{ij} + \beta_{4} * LP_{i} + \beta_{5} * LP_{j} * epi^{1}_{ij}$$
$$+ \beta_{6} * LP_{j} * epi^{2}_{ij} + \beta_{7} * LP_{j} * epi^{3}_{ij} + [u_{o}{}_{ij} + u_{1}{}_{i} * epi_{ij}]$$

Three notes:
1. A_{ij} = the proportion of an activity of writer *i* (*j* = 1, 2, …, 20) during task *j* (*j* = 1, 2, …, 8). LP = individual language proficiency scores. Before entering the language proficiency scores into the model, they were converted into *z* scores.
2. Note that epi^{0} = 1.
3. $Episode^{2}$, $episode^{3}$, LP*$episode^{2}$ and LP*$episode^{3}$ were not included as a term for all of the seven activities. As explained in the 'Method' section, inclusion of these terms depends on whether lower-order terms have reached significance and whether the term itself makes a significant contribution to the model.

APPENDIX I (chapter 4, p. 70; chapter 4, p. 79; chapter 6, p. 125) Parameter estimates (se) for the average occurrence (in proportions) of seven cognitive activities (β_0, β_1, β_2, and β_3), as well as the variances due to writer (S^2_{u0i} and S^2_{u1i}) and due to task (S^2_{u0j}). Only significant estimates (p<.05) are presented.

		Fixed parameters			
		Average occurrence			
		β_0 *epi^0	β_1 *epi^1	β_2 *epi^2	β_3 *epi^3
Reading the Assignment	L1	.10 (.01)	-.02 (.002)	-	-
	L2	.23 (.03)	-.17 (.04)	.05 (0.01)	-.005 (.001)
Process Planning	L1	.12 (.02)	-.08 (.02)	.02 (0.01)	-.002 (.001)
	L2	.07 (.009)	-.03 (.004)	.005 (.0007)	-
Content Planning	L1	.04 (.008)	-.005 (.001)	-	-
	L2	.05 (.01)	-.006 (.002)	-	-
Formulating	L1	.27 (.02)	.09 (.01)	-.01 (.002)	-
	L2	.21 (.03)	.11 (.01)	-.02 (.002)	-
Reading own Text	L1	.02 (.004)	-	-	-
	L2	.01 (.004)	-	-	-
Evaluating Own Text	L1	-.01 (.007)	.02 (.01)	-.008 (.004)	.001 (.0004)
	L2	.00007 (.002)[12]	.002 (.0008)	-	-
Revising	L1	.05 (.007)	-	-	-
	L2	.04 (.006)	-	-	-

		Random parameters		
		Writer		Task
		S^2_{u0i}	S^2_{u1i}	S^2_{u0j}
Reading the Assignment	L1	.002 (.001)	-	-
	L2	.004 (.001)	.0001 (.00006)	-
Process Planning	L1	.0001 (.0004)	-	-
	L2	.001 (.0004)	.00004 (.00002)	-
Content Planning	L1	.0009 (.0004)	.00003 (.00001)	-
	L2	.003 (.001)	.00006 (.00003)	-
Formulating	L1	.005 (.002)	.0003 (.0001)	-
	L2	.009 (.003)	.0003 (.0001)	-
Reading own Text	L1	.0003 (.0001)	-	-
	L2	.0002 (.00008)	-	-
Evaluating Own Text	L1	.00002 (.0000)	.00002 (.0000)	-
	L2	.000002 (.0000)	.000009 (.0000)	-
Revising	L1	.0005 (.0002)	-	.0002 (.0001)
	L2	.0004 (.0001)	-	.0001 (.00007)

[12] This parameter reached significance before epi^1 was entered into the model.

APPENDIX J (chapter 4, p. 75)

Parameter estimates (se) for variations in occurrence (in proportions) due to Language Proficiency (*LP*) scores. Significant effects (p<.05) are marked with *.

	L1		L2	
	β_4*LP	$\beta_5*LP*\mathrm{epi}^1$	β_4*LP	$\beta_5*LP*\mathrm{epi}^1$
Reading the Assignment	.006 (.011)	.00001 (.002)	.002 (.015)	-.001 (.003)
Process Planning	.002 (.008)	-.001 (.001)	-.018 (.007)*	.0033 (.0015)*
Content Planning	-.009 (.007)	.003 (.001)	-.0224 (.0115)	.004 (.002)*
Formulating	-.013 (.018)	.006 (.004)	.02 (.02)	-.002 (.005)
Reading Own Text	-.00004 (.004)	-.0000001 (.0004)	-.001 (.003)	-.00004 (.0005)
Evaluating Own Text	.004 (.002)*	-.002 (.001)*	-.0009 (.001)	-.0002 (.0008)
Revising	.000002 (.005)	.00008 (.001)	.005 (.004)	.00004 (.001)

Note that the L1 and L2 language proficiency tests are not directly comparable. For instance: in the L2 language proficiency tests, the L1 translation of the looked-for word is given behind brackets, which is not the case in the L1 language proficiency tests. Although this was not the aim of this study, it should be noted that the regression weights (β4 and β5) in L1 and L2 are not directly comparable either.

APPENDIX K (chapter 5, p. 94)

Let A_{ij} be the proportion of an activity of writer j (j = 1, 2, …, 20) at episode i (i = 1, 2, …, 5), and P_j and R_j the planner and reviser scores of this writer. The model used to analyse the data can be written as:

$$Logit(A_{ij}) = \beta_o * epi^0{}_{ij} + \beta_1 * epi^1{}_{ij} + \beta_2 * epi^2{}_{ij} + \beta_3 * P_j + \beta_4 * epi^1{}_{ij} * P_j$$

$$+ \beta_5 * R_j + \beta_6 * epi^1{}_{ij} * R_j + [u_{0j} + u_{1j} + epi_{ij}]$$

Three notes:
1. Note that $\text{epi}^0 = 1$.
2. Remember: *Logit (Y) = Ln (Y / 1 –Y).2.* The dependent variables are proportions. As both the means and their variances are estimated (i.e. the variance of the residuals), in effect the proportion for each individual is estimated. Therefore, the intra-individal variance is not estimated. Only the variance of the between-writer residuals (i.e. u_{0j} and u_{1j}) is estimated.
3. *Episode*^{2} was not included as a term for all of the six activities. This was, as explained in the 'Method' section dependent on whether lower-order terms have reached significance and whether the term itself makes a significant contribution to the model.

APPENDIX L (chapter 5, p. 96)

Parameter estimates for the average distributions for all of six (meta)cognitive activities (β_0, β_1, and, β_2), as well as the variances (S^2_{u0j} and S^2_{u1}) and the respective standard errors (between brackets). Note that 'episode' was re-scaled around the mean episode (= episode 3).

If the ratio of a fixed parameter exceeds 1.96 times its associated standard error, it is assumed to contribute significantly to the fit of the model. For random parameters, a more relaxed criterion for significance is used. After all, variances cannot take negative values. Therefore, we are dealing with a one-sided hypothesis being tested. A variance estimate contributes significantly if it exceeds 1.65 times the associated standard error. So, in effect a 5% significance level is in operation for both fixed and random parameters.

All estimates in this table are significant.

	Fixed parameters			Random parameters	
Activity	β_0 *epi^0	β_1 *epi^1	β_2 *epi^2	S^2_{u0j}	S^2_{u1j}
Reading assignment	-2.91 (.09)	-.31 (.04)	.04 (.02)	.47 (.09)	.07 (.02)
Planning	-3.00 (.11)	.25 (.03)	.04 (.02)	.77 (.14)	.03 (.01)
Text production	-0.37 (.04)	.01 (.00)	-.06 (.01)	.06 (.01)	.01 (.00)
Reading own text	-3.69 (.12)	.28 (.04)	.06 (.02)	.78 (.15)	.09 (.02)
Evalating	-5.38 (.17)	.38 (.09)	.09 (.05)	.70 (.22)	.28 (.09)
Revising	-8.16 (.58)	.25 (.02)	.26 (.08)	4.22 (1.67)	1.62 (.62)

These estimates are expressed in logits. How to read and interpret them is described below. The activity 'reading the assignment' is used as an example.

As the 'episode' variable was centred, episode 1 became episode -2, episode 2 became episode -1, episode 3 became episode 0, et cetera. For reading the assignment, then, the (mean) logit in the first episode equals [-2.91*-2^0+ -0.31*-2^1 + 0.4* -2^2 =] -2.13. When converted to proportions this equals [expo(-2.13)/(1+expo(-2.13)) =] 0.10 (see also Figure 3).

The variance between writers is also significant for reading the assignment. The intercept variance equals 0.47. Hence, in episode 3 (re-scaled as episode 0) the 80% (z score = 1.28) confidence intervals for the observed proportions varies from [logit(-2.91 – √0.47*1.28) =] 0.05 to [logit (-2.91 + √0.47*1.28) =] 0.22. Writers also differ from each other with regard to the linear change of reading the assignment. This means that the 80% confidence interval for the linear change varies from [logit (-.31- √.07 *1.28) =] -0.61 to [logit (-.31+ √.07 *1.28) =] 0.26. Whereas for most students there is a clear decrease in the proportions of reading the assignment, some students show an increase in this activity during writing.

The parameter estimates in logits are somewhat hard to interpret. Therefore the mean changes over the five episodes are presented in proportions in Figure 3. Episode numbers were, for the sake of ease of interpretation, also reconverted into their original numbers.

APPENDIX M (chapter 5, p. 99)

Variation in distributions due to Writing Questionnaire Scores. Parameter estimates and standard errors (between brackets) in logits. P: Planners core; R: Reviser score.

		Fixed parameters		
Activity	β_3 * P_j	β_4 * P_j *epi^1_j	β_5 *R_j	β_6 * Rj *epi^1_j
Reading the assignment	-.227 (.09)	-.012 (.038)	.105 (.086)	.074 (.037)
Planning	-.150 (.065)	-.100 (.030)	.061 (.031)	-.014 (.009)
Text production	.087 (0.031)	.002 (.014)	.040 (.031)	-.006 (.015)
Reading own text	-.082 (.104)	.060 (.047)	-.252 (.106)	.009 (.048)
Evaluating	.172 (.128)	.044 (.077)	.021 (.138)	.003 (.084)
Revising	1.181 (.305)	.189 (.232)	.535 (.472)	-.055 (.327)
		Random parameters		
		β_0	β_1	
Reading the assignment	β_0	.428** (.081)	--	
	β_1	.115** (.030)	.063** (.016)	
		β_0	β_1	
Planning	β_0	.789** (.140)	--	
	β_1	.043** (.027)	.028** (.009)	
		β_0	β_1	
Text production	β_0	.052** (.011)	--	
	β_1	.001 (.004)	.008** (.002)	
		β_0	β_1	
Reading own text	β_0	.601** (.119)	--	
	β_1	-.020 (.038)	.092** (.024)	
		β_0	β_1	
Evaluating own text	β_0	.613** (.202)	--	
	β_1	.066 (.087)	.199** (.074)	
		β_0	β_1	
Revising	β_0	1.320** (.968)	--	
	β_1	-.442 (.577)	1.301** (.623)	

APPENDIX N (chapter 6, p. 112)

Parameter estimates for the models explaining text quality with episodes, language proficiency (*z scores: zLP*), and interactions between episodes and language proficiency. Significant estimates (p < .05) are marked with *.

Reading the Assignment

Predictor	**L1**	**L2**	**Predictor**	**L1**	**L2**
Intercept	112.19*		zLP (L1 or L2)	21.00*	5.09
Epi 1	0.45	-0.59*	Epi 1 * zLP	-0.52	-0.42
Epi 2	1.40*	-0.77	Epi 2 * zLP	-1.19*	-2.21*
Epi 3	0.50	-0.61	Epi 3 * zLP	0.09	-0.05
Epi 4	0.75	-1.16	Epi 4 * zLP	-1.85*	-1.83*
Epi 5	0.56	2.85*	Epi 5 * zLP	-1.24	2.15

Process Planning

Predictor	**L1**	**L2**	**Predictor**	**L1**	**L2**
Intercept	119.01*		zLP (L1 or L2)	9.54*	-1.87
Epi 1	0.40	-2.11*	Epi 1 * zLP	-0.11	0.67
Epi 2	0.22	0.73	Epi 2 * zLP	1.42*	4.49*
Epi 3	2.76*	-0.12	Epi 3 * zLP	-1.93*	-4.76*
Epi 4	-2.29*	-2.67*	Epi 4 * zLP	-0.70	-0.63
Epi 5	2.18*	-2.12*	Epi 5 * zLP	-2.02	-3.47*

Content Planning

Predictor	**L1**	**L2**	**Predictor**	**L1**	**L2**
Intercept	112.40*		zLP (L1 or L2)	4.50	-7.00*
Epi 1	2.00*	0.16	Epi 1 * zLP	-1.65*	-0.44
Epi 2	-1.99	0.51	Epi 2 * zLP	-0.49	-0.63
Epi 3	3.23*	-0.78	Epi 3 * zLP	2.61	1.52
Epi 4	0.97	-2.37*	Epi 4 * zLP	-1.46	-1.78
Epi 5	-0.09	-0.14	Epi 5 * zLP	-0.01	-0.11

Formulating

Predictor	**L1**	**L2**	**Predictor**	**L1**	**L2**
Intercept	111.14*		zLP (L1 or L2)	-1.86	7.75
Epi 1	0.23	-0.11	Epi 1 * zLP	0.43	0.04
Epi 2	-0.01	-0.58	Epi 2 * zLP	-0.32	-0.89*
Epi 3	0.55	0.24	Epi 3 * zLP	0.68	1.43*
Epi 4	0.44	1.02*	Epi 4 * zLP	-0.25	0.22
Epi 5	-0.64*	-1.08*	Epi 5 * zLP	-0.39	-0.87

Reading Own Text					
Predictor	**L1**	**L2**	**Predictor**	**L1**	**L2**
Intercept	112.39*		zLP (L1 or L2)	7.96	21.46*
Epi 1	1.80	-2.44	Epi 1 * zLP	-2.04	-2.04
Epi 2	0.60	-1.10	Epi 2 * zLP	-0.39	-2.65
Epi 3	-0.64	-0.79	Epi 3 * zLP	-1.13	-1.91
Epi 4	1.40	-0.98	Epi 4 * zLP	2.69*	0.85
Epi 5	1.39*	-0.62	Epi 5 * zLP	-0.19	-1.89*

Evaluating Own Text					
Predictor	**L1**	**L2**	**Predictor**	**L1**	**L2**
Intercept	114.72*		zLP (L1 or L2)	4.53	3.47
Epi 1	-9.36	-0.04	Epi 1 * zLP	-6.57	0.96
Epi 2	6.23*	-5.99*	Epi 2 * zLP	-0.07	-0.75
Epi 3	-2.95	5.20	Epi 3 * zLP	1.76	2.39
Epi 4	5.04	-3.13 *	Epi 4 * zLP	4.95	3.30
Epi 5	4.84*	-4.95*	Epi 5 * zLP	1.75	1.29

Revising					
Predictor	**L1**	**L2**	**Predictor**	**L1**	**L2**
Intercept	101.97*		zLP (L1 or L2)	-8.88	5.56
Epi 1	0.50	-0.05	Epi 1 * zLP	-0.70	-0.53
Epi 2	0.56	-0.13	Epi 2 * zLP	-0.04	0.48
Epi 3	1.72*	-0.41	Epi 3 * zLP	-1.57*	-2.86*
Epi 4	1.04*	-0.50	Epi 4 * zLP	3.58*	4.11*
Epi 5	0.18	0.00	Epi 5 * zLP	0.03	-0.16

SCHRIJVEN IN MOEDERTAAL EN VREEMDE TAAL

(Nederlandstalige samenvatting van dit proefschrift)

Dit proefschrift gaat over de schrijfvaardigheid van leerlingen in het voortgezet onderwijs. Hoewel het is gebaseerd op een veelheid aan eerder onderzoek, vormen twee onderzoeksstromen het fundament onder het beschreven onderzoek. Zeer belangrijk waren, ten eerste, de theoretische modellen van Flower en Hayes (1980), Hayes en Flower (1980) en Hayes (1996), waarin (voornamelijk) cognitieve aspecten van het schrijfproces worden beschreven. Ten tweede is het onderzoek in dit proefschrift gebaseerd op het 'probabilistische model' van schrijfprocessen gepresenteerd door Rijlaarsdam en Van den Bergh (1996) en Van den Bergh en Rijlaarsdam (1996). Deze onderzoeksstromen worden hieronder kort toegelicht.

Flower en Hayes (1980), Hayes en Flower (1980) en Hayes (1996): cognitieve schrijfprocesmodellen

Flower en Hayes (1980) en Hayes en Flower (1980) zetten met hun intussen welbekende cognitieve model van schrijfprocessen de complexiteit van de activiteit 'schrijven' op de kaart. Dit model demonstreerde dat schrijven een vaardigheid is met veel facetten, die een beroep doet op een veelheid aan subvaardigheden en kennis. Hayes (1996) presenteerde een herziene versie van het model uit 1980. Het model bestaat uit twee componenten: de *taakomgeving* en het *individu*. De taakomgeving bevat behalve de fysieke taakomgeving (de eigen-tekst-tot-zover en de opdracht), ook de sociale context waarin geschreven wordt. De individuele component in het model bevat vier subcomponenten, namelijk: het langetermijngeheugen (met daarin kennis van onderwerp, publiek en genre, procedurele kennis, en taalkennis), het schrijfproces (bestaand uit diverse cognitieve activiteiten), het werkgeheugen, en motivatie. De complexiteit van schrijven is gelegen in het feit dat er tijdens het schrijven aandacht moet worden besteed aan alle beschreven subcomponenten, soms tegelijkertijd.

Rijlaarsdam & Van den Bergh (1996) en Van den Bergh & Rijlaarsdam (1996): temporele benadering van schrijfprocessen

Rijlaarsdam en Van den Bergh (1996) en Van den Bergh en Rijlaarsdam (1996) presenteerden een zogenaamd 'probabilistisch model van schrijfprocessen'. Een cruciaal aspect van dit model is de gedachte dat de functie van een cognitieve activiteit (zoals lezen, plannen, formuleren, reviseren) afhangt van de context waarin deze wordt uitgevoerd. Bijvoorbeeld: het lezen van de opdracht heeft aan het begin van het schrijfproces een andere functie dan aan het einde van het schrijfproces. Aan het begin is het waarschijnlijk bedoeld om een idee te krijgen van waar de tekst over moet gaan, wie het publiek is, wat de te bereiken doelen zijn, enzovoorts. Tegen het einde van het schrijfproces is het lezen van de opdracht waarschijnlijk onderdeel van evaluerende activiteiten: voldoet de geproduceerde tekst aan de in de opdracht gestelde eisen?

Deze ideeën worden empirisch ondersteund. De onderzoekers lieten bijvoorbeeld zien dat de effectiviteit van cognitieve activiteiten verschillend is op verschillende momenten gedurende het schrijfproces. Het optreden van de activiteit 'structureren', bijvoorbeeld, was aan het begin van het schrijfproces effectiever dan aan het einde van het schrijfproces. Dat wil zeggen, structureren correleerde aan het begin van het schrijfproces sterker met tekstkwaliteit dan aan het einde van het schrijfproces (Rijlaarsdam & Van den Bergh, 1996). De onderzoekers benadrukten daarom het belang van temporele analyses van hoe cognitieve activiteiten verdeeld zijn over het schrijfproces. Bij bestudering van schrijfprocessen moeten onderzoekers dus rekening houden met de momenten waarop cognitieve activiteiten zich gedurende taakuitvoering voordoen.

Deze temporele benadering is in diverse studies gebruikt en zinvol gebleken (Breetvelt, Van den Bergh & Rijlaarsdam, 1994; Leijten & Van Waes, 2006; Olive, Kellogg & Piolat, 2008; Roca de Larios, Marín & Murphy, 2011; Rijlaarsdam & Van den Bergh, 1996; Rijlaarsdam et al., 2011; Van den Bergh & Rijlaarsdam, 1996; Van den Bergh et al., 2009; Van der Hoeven, 1997; Van Weijen, Van den Bergh, Rijlaarsdam & Sanders, 2008).

Schrijven in eerste taal en tweede taal

In een wereld die steeds meer globaliseert wordt het steeds belangrijker dat men leert om zich uit te drukken in andere talen dan enkel de moedertaal (of eerste taal; T1). In het Nederlands voortgezet onderwijs wordt een aantal vreemde talen onderwezen. Doorgaans legt men de meeste nadruk gelegd op de verwerving van het Engels. Het is de taal die in het basisonderwijs al wordt aangeboden, en de enige taal waarin vrijwel alle leerlingen eindexamen (moeten) doen in het voortgezet onderwijs. Dit drukt het belang uit dat kennelijk aan het Engels wordt gehecht.

Veel kinderen komen, bijvoorbeeld via de media, al vanaf jonge leeftijd in aanraking met het Engels. Het Engels wordt dus formeel onderwezen als 'vreemde taal', maar daarnaast wordt het ook op een meer natuurlijke wijze verworven. Daarom zou men kunnen stellen dat het Engels in Nederland voor veel mensen haast als een tweede taal (T2) fungeert.

Hoewel leerlingen in het voortgezet onderwijs zich vaak al redelijk kunnen uitdrukken in het Engels, is hun Engelse taalvaardigheid (T2) toch meestal van een substantieel lager niveau dan hun Nederlandse taalvaardigheid (T1). Dit taalvaardigheidsverschil wordt meestal gegeven als verklaring voor het vaak geobserveerde kwaliteitsverschil tussen schrijfvaardigheid in T1 en T2 (Sasaki & Hirose, 1996; Schoonen et al., 2003). Een mindere taalvaardigheid (in T2 ten opzichte van T1) kan de kwaliteit van schrijven op twee manieren beïnvloeden. Ten eerste beperkt de lagere taalvaardigheid het vermogen van leerlingen om hun ideeën (correct) onder woorden te brengen. Ten tweede wordt aangenomen dat taalvaardigheidsproblemen het werkgeheugen zodanig belasten, dat er minder werkgeheugencapaciteit overblijft voor conceptuele en regulerende activiteiten (zoals structureren en het 'managen' van het schrijfproces) en/of voor het transfereren van T1-schrijfstrategieën naar T2-schrijftaken.

Beschrijving van het onderhavige onderzoek: dataverzameling

Het belangrijkste doel van het hier beschreven onderzoek was om inzicht te krijgen in (de oorzaken van) het kwaliteitsverschil tussen T1- en T2-schrijven. Er zijn op empirische wijze data verzameld om een vergelijking tussen T1- en T2-schrijven mogelijk te maken.

Er namen twintig leerlingen uit de derde klas van het voortgezet onderwijs (stroming vwo) deel aan het onderzoek. Zij schreven ieder vier korte argumentatieve opstellen in het Nederlands (T1) en vier korte argumentatieve opstellen in het Engels (T2). T1- en T2-tekstkwaliteit werden vergeleken, maar ook T1- en T2-schrijfprocessen. Het is essentieel dat leerlingen per taal meerdere schrijfopdrachten uitvoeren. Immers, als er slechts één schrijfopdracht per taal wordt gebruikt, is het onmogelijk vast te stellen of eventuele verschillen die na vergelijking worden gevonden, veroorzaakt zijn door verschillende talen of door verschillende schrijftaken (Van den Bergh et al., 2009; Van Weijen, 2009).

De kwaliteit van iedere tekst werd beoordeeld door drie beoordelaars, die onafhankelijk van elkaar tot hun oordeel kwamen. In totaal namen er acht beoordelaars deel aan het onderzoek. De schrijfprocessen werden geanalyseerd in termen van de volgende cognitieve activiteiten: lezen van de opdracht, plannen van het proces, plannen van inhoud, formuleren, lezen van de eigen tekst, evalueren van eigen tekst, en reviseren. De schrijfprocessen werden geregistreerd door middel van de hardopdenkmethode gecombineerd met toetsaanslagregistratie (Leijten & Van Waes, 2006). Toetsaanslagregistratie heeft een betrouwbaarheidsvoordeel: de registratie geschiedt automatisch, zonder in te breken in het schrijfproces. Daarnaast verschaft de registratie van toetsaanslagen zeer gedetailleerde informatie over revisies en over pauzes in het schrijfproces. De hardopdenkmethode is juist geschikt om informatie te verkrijgen over meer conceptuele cognitieve activiteiten, zoals plannen en evalueren. Het verkrijgen van informatie over denkprocessen kent beperkingen, maar door toetsaanslagregistratie te combineren met de hardopdenkmethode, worden er data verkregen die zo betrouwbaar en volledig zijn als mogelijk is.

De leerlingen maakten ook taalvaardigheidstoetsen Nederlands en Engels, en vulden een vragenlijst in waarin zij hun eigen schrijfstijl rapporteerden.

Belangrijkste uitkomsten

Tekstkwaliteit

Hoewel men het erover eens is dat T2-teksten gemiddeld van lager niveau zijn dan T1-teksten, is dit kwaliteitsverschil nog nooit gekwantificeerd. In eerder onderzoek heeft men wel geïsoleerde kenmerken van T1- en T2-teksten vergeleken. Silva (1993) concludeerde bijvoorbeeld dat T2-teksten gemiddeld korter zijn, meer taalfouten bevatten, minder samenhangende argumentatie bevatten, en minder gericht zijn op de lezer. Tekstscores die de kwaliteit van teksten in hun geheel uitdrukken waren echter nooit vergelijkbaar. Om een dergelijke vergelijking mogelijk te maken moeten de T1- en T2-tekstscores namelijk worden uitgedrukt op één en dezelfde schaal. Zo een schaal bestond tot op heden niet. Beoordelaars bleken (na statistische toetsing) bijvoorbeeld niet even streng zijn voor T1- en T2-teksten. Men moet daarom aannemen dat beoordelingen van T1- en T2-teksten doorgaans op verschillende schalen zijn uitgedrukt, en dus niet vergelijkbaar zijn.

In dit proefschrift werd een procedure getest waarmee directe vergelijking van T1- en T2-tekstscores mogelijk bleek. Twee kenmerken definiëren deze procedure. Ten eerste zijn de beoordelaars tweetalig (T1/T2) of beheersen zij zowel T1 als T2 praktisch op het niveau van een moedertaalspreker, om zo de kans te vergroten dat zij dezelfde maatstaven toepassen op zowel T1- als T2-teksten. Ten tweede worden de beoordelingen uitgevoerd met ankeropstellen (d.w.z. modelopstellen die de gemiddelde kwaliteit weergeven) in zowel T1 als T2. Omdat de beoordelingen met T1- en T2-ankeropstellen in statistische zin parallelle toetsen bleken, is er geen reden om aan te nemen dat de beoordelaars andere maatstaven gehanteerd hebben ten opzichte van T1 en T2. Omdat de beoordelingen daarnaast voldoende betrouwbaar bleken, konden T1- en T2-tekstscores worden vergeleken. Binnen talen bleken er grote verschillen in tekstscores te bestaan ten gevolge van verschillende schrijftaken. Het toegevoegde effect van taal is echter vijf maal zo groot. Dus hoewel een deel van de verschillen in tekstscores wordt verklaard doordat er werd geschreven naar aanleiding van een andere schrijfopdracht, is het toegevoegde, kwaliteitsverminderende, effect van schrijven in T2 op de tekstscores nog veel groter. De kwaliteit van T2-teksten was zoveel lager dan die van T1-teksten, dat er slechts een zeer kleine overlap bestond tussen de beste T2-teksten en de slechtste T1-teksten. Dit betekent dat veel van de matige T1-teksten toch nog van hogere kwaliteit waren dan goede T2-teksten.

Een andere interessante uitkomst is dat individuele beoordelaars niet voldoende betrouwbaar waren, maar dat dit probleem wordt opgeheven door de beoordelingen van drie beoordelaars samen te voegen tot jurybeoordelingen. Door teksten te laten beoordelen door meer beoordelaars, is de kans veel groter dat een tekstscore de werkelijke kwaliteit van die tekst weerspiegelt.

Schrijfprocessen

In dit proefschrift werd gedemonstreerd dat effectieve schrijfprocessen in T2 verschillen van effectieve schrijfprocessen in T1. Veel cognitieve activiteiten bleken op andere momenten gedurende het schrijfproces relevant voor tekstkwaliteit. In T1 is het bijvoorbeeld aan het begin en in het midden van het schrijfproces essentieel voor de kwaliteit van de resulterende tekst dat men zich veel bezighoudt met inhoud plannen, terwijl de frequentie van inhoud plannen in T2 op die momenten niet gerelateerd is aan tekstkwaliteit. Het maakt aan het begin en in het midden van het T2-schrijfproces dus niet uit of men veel of weinig plant. In T2 is de mate waarin er inhoud gepland wordt juist tegen het einde van het schrijfproces relevant voor tekstkwaliteit (hoe meer inhoud er tegen het einde gegenereerd wordt, hoe minder de kwaliteit van de tekst). Met andere woorden, leerlingen moeten hun aandacht tijdens T2-schrijven anders over het proces van taakuitvoering verdelen dan tijdens T1-schrijven.

Wanneer cognitieve activiteiten in T1 en T2 wel op dezelfde momenten in het schrijfproces relevant zijn voor tekstkwaliteit, is de relatie met tekstkwaliteit in T1 en T2 vaak tegengesteld. Het evalueren van de geschreven tekst is aan het einde van T1-processen bijvoorbeeld positief gerelateerd aan tekstkwaliteit (hoe meer evalueren, hoe beter de tekst), maar aan het einde van T2-processen negatief. Dus ook al wordt er in T2, naar T1-maatstaven, op het juiste moment geëvalueerd, de uitvoering van deze evaluaties is kennelijk niet gunstig voor de kwaliteit van de resulterende (T2-)tekst. Deze uitkomst roept vragen op. Hoe kan het dat het uitvoeren van dezelfde activiteit, op hetzelfde moment, kenmerkend is voor succesvol schrijven in T1, maar voor mindere schrijfvaardigheid in T2?

Additionele analyses (zie hoofdstuk 6) laten zien dat deze tegenstelling verdwijnt wanneer taalvaardigheid wordt meegenomen als verklarende factor. De T1- en T2-relaties tonen gelijke richtingen voor leerlingen die meer taalvaardig zijn in T2. Reviseren aan het einde van het schrijfproces is bij een hogere taalvaardigheid bijvoorbeeld gunstig voor tekstkwaliteit in zowel T1 als T2, en ongunstig voor tekstkwaliteit bij een lagere taalvaardigheid in zowel T1 als T2. Deze additionele analyses suggereren dat T1- en T2-schrijven bij voldoende taalvaardigheid tamelijk gelijke eisen stellen in termen van de mate waarin cognitieve activiteiten op bepaalde momenten worden toegepast.

Suggesties voor het onderwijs

Hoewel het onderzoek beschreven in dit proefschrift correlationeel van aard is, en oorzakelijke verbanden tussen onderzochte constructen strikt genomen niet kunnen worden vastgesteld, en hoewel dit onderzoek niet als direct doel had om onderwijsstrategieën voor de ontwikkeling van schrijfvaardigheid te testen, zijn niettemin een aantal bevindingen van belang voor de onderwijspraktijk.

Ten eerste onderstreept dit onderzoek het belang van beoordeling van schrijfvaardigheid met meerdere taken en meerdere beoordelaars. Verschillende taken roepen, ook al verschillen ze slechts in termen van het onderwerp waarover geschreven moet worden, flinke fluctuaties in tekstscores op. Dit is bepaald geen nieuw inzicht (Coffman, 1966; Van den Bergh, 1988b; Wesdorp, 1974). Van den Bergh (1988b) beschreef bijvoorbeeld al dat beoordelingen van schrijfvaardigheid op basis van slechts één taak eigenlijk toetsen zijn met slechts één item. Niettemin is het nog vrij gangbaar om leerlingen te becijferen op basis van hun schrijfprestaties gedurende één taak. Dit onderzoek demonstreert andermaal dat een dergelijke gang van zaken grote betrouwbaarheidsproblemen kent. Op eenzelfde wijze zijn oordelen over schijfvaardigheid door slechts één beoordelaar zeer waarschijnlijk onbetrouwbaar.

Ten tweede suggereert dit onderzoek dat het zeer effectief kan zijn als leerlingen, zeker bij een lagere T2-taalvaardigheid, hun schrijfprocessen anders inrichten tijdens T2-schrijftaken dan tijdens T1-schrijftaken. Met andere woorden, het is niet altijd goed als leerlingen zich gedurende T2-schrijven vasthouden aan T1-schrijfstrategieën. Hun verminderde T2-taalvaardigheid kan ervoor zorgen dat zij hun T1-werkwijze niet met voldoende kwaliteit kunnen toepassen tijdens T2-schrijftaken.

CURRICULUM VITAE

Marion Tillema was born on March 31st 1981 in Groningen, The Netherlands. She obtained her *atheneum* diploma from the *Wessel Gansfortcollege* in Groningen in 1999. In 2003, she received an MA degree in Linguistics from Utrecht University (specializing in Second Language Learning). After graduating, she worked as an educational author for Malmberg Publishers, and as a PhD candidate at the Utrecht Institute of Linguistics – OTS. This dissertation is the result of the research she carried out during her PhD period.