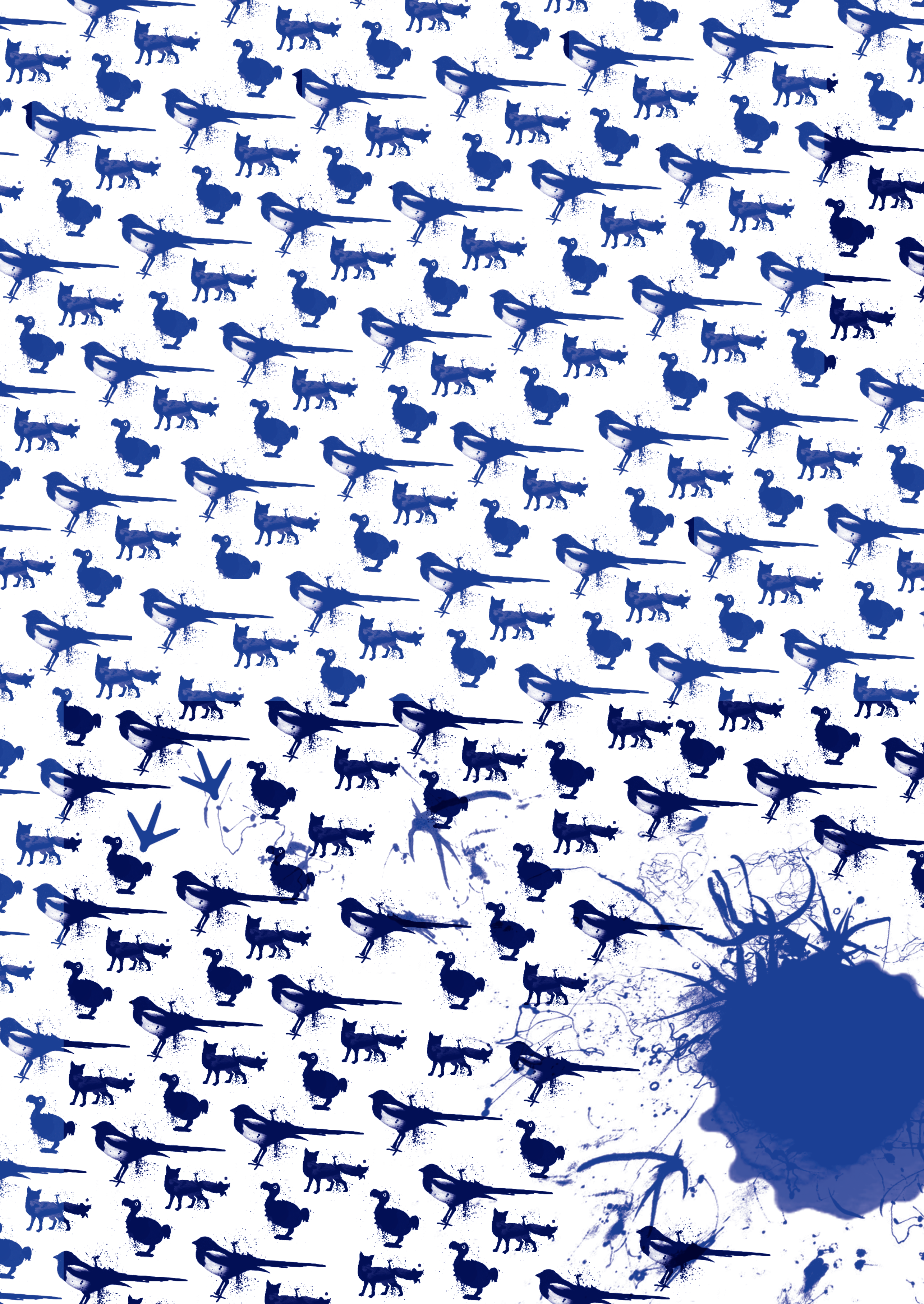
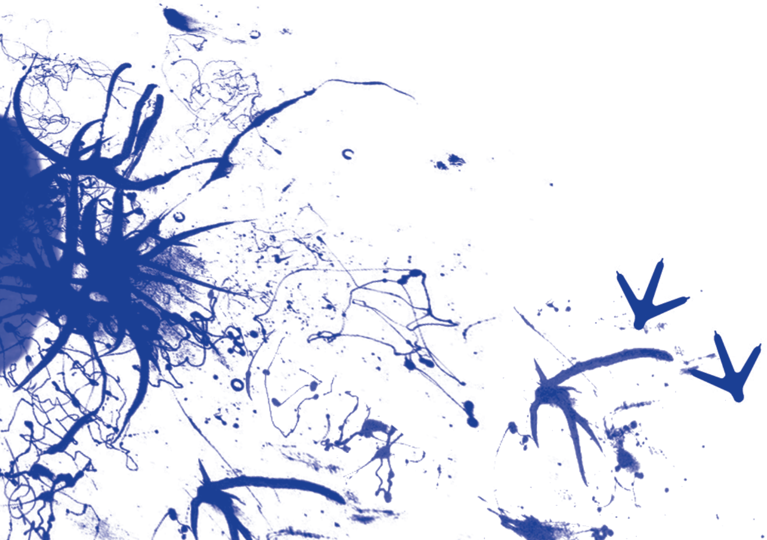
BRINGING WRITING RESEARCH INTO THE CLASSROOM

The effectiveness of Tekster, a newly developed writing program for elementary students

Renske Bouwer & Monica Koster





The research reported here was carried out at the

Utrecht Institute of Linguistics OTS

in the context of the research school

ico

Interuniversity Centre for Educational Research

ISBN 978-94-6103-052-8

Design: Rémi Bouwer
Printed by: Pantheon Drukkers

BRINGING WRITING RESEARCH INTO THE CLASSROOM

The effectiveness of Tekster, a newly developed writing program
for elementary students

SCHRIJFONDERZOEK NAAR HET KLASLOKAAL

De effecten van Tekster, een nieuw ontwikkeld lesprogramma
voor leerlingen in het basisonderwijs

(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 vrijdag 2 september 2016

te 2.30 uur door

Monica Patricia Koster

geboren op 29 mei 1967
te Den Helder

te 3.15 uur door

Ilse Renske Bouwer

geboren op 20 april 1984
te Purmerend

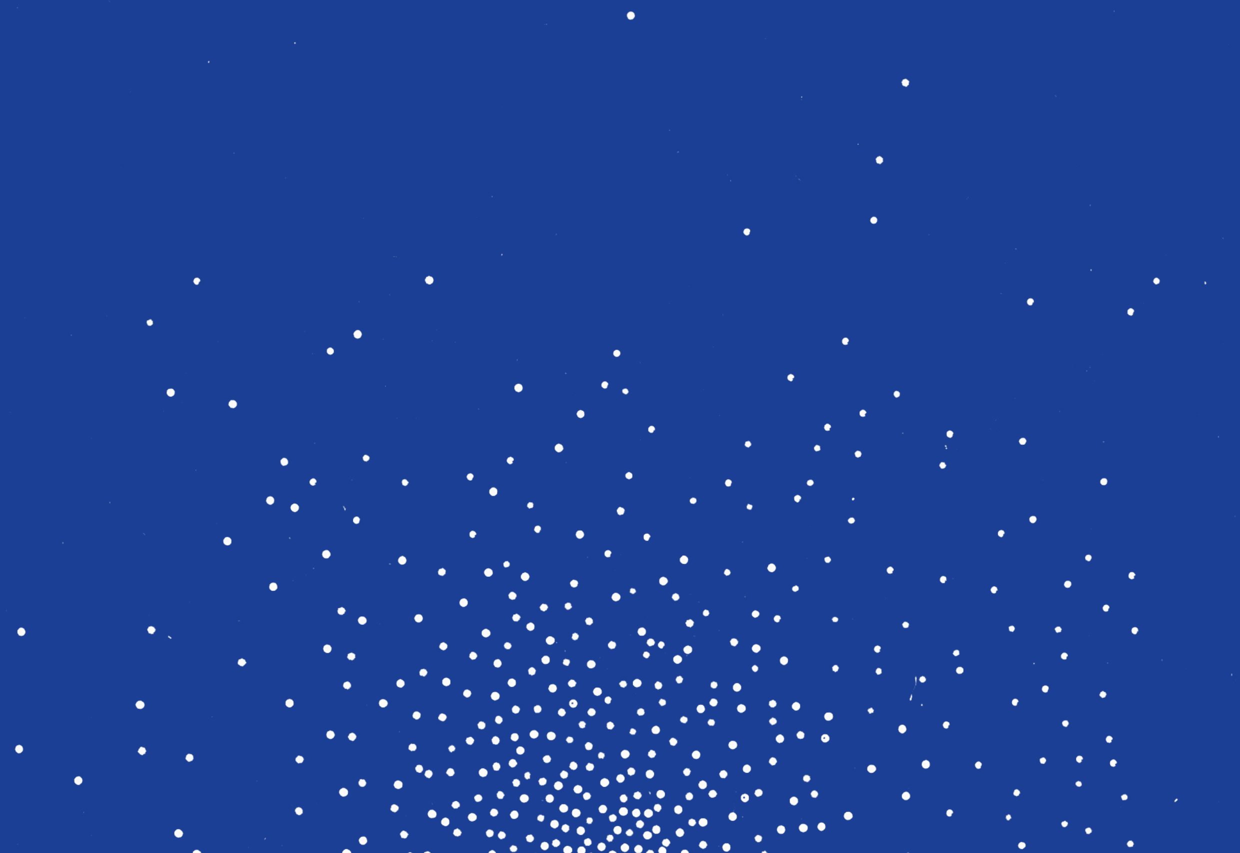
Promotoren:	Prof.dr. H.H. van den Bergh
	Prof.dr. T.J.M. Sanders
Copromotor:	Dr. A. Béguin

Special thanks to Vos, Dodo & Ekster

TABLE OF CONTENTS

Chapter 1

INTRODUCTION



IMPORTANCE OF LEARNING TO WRITE

Nowadays, writing is more important than ever. Where until the early twentieth century writing skills used to belong to an elite, today writing skills are important for everyone (Brandt, 1995). During the last century our society has shifted from a mainly agricultural and manufacturing economy to a knowledge economy, where people earn their living exchanging and processing information. This shift has brought about tremendous changes in everyday life and has resulted in an overwhelming accumulation of written information. To deal with this information overload it is increasingly important to develop adequate literacy skills (Brandt, 1995). On top of that, the rapid revolution in communication technology from the start of the 21st century has profoundly changed the ways we communicate. Through all our digital devices we are able to interact 24/7. Whereas it seems that most of this communication is visual, through images (Instagram) and videos (YouTube), the majority of communication is still in written form, e.g., e-mail, social media, blogs, websites, and reports.

Writing is a foundational skill that can boost comprehension and achievement, as writing enables students to organize their thinking and think critically (Graham, 2013). Writing proficiency is associated with academic and occupational success (Graham & Perin, 2007). Hence, to successfully participate in society, it is of vital importance that children learn to write well already at a very young age. It is a common misconception that writing is an innate ability, like singing or drawing, which you are either good at or not. Another misconception about writing is that 'practice makes perfect'. Of course a writer needs practice to improve, but he also needs instruction and feedback on how he can improve. Contrary to oral language skills, which are acquired naturally at home during the early years of development through exposure and practice in daily life, writing skills are acquired through explicit instruction and guided practice. When students are not given sufficient time and explicit instruction in writing, they can never meet the demands that are so important for full participation in 21st century daily life. According to Graham (2013, p. 4): "Nowhere is this more important than in the early grades, where the foundation for future achievement is established". This means that we must make writing instruction a priority in elementary classrooms.

CURRENT STATE OF WRITING EDUCATION

In the last decade worldwide concerns are raised on the level of students' writing proficiency, in the US (cf. National Center for Education Statistics, 2012), as well as in Western Europe (cf. Neumann 2012; Ofsted, 2012). In the Netherlands, two large national assessment studies (Krom, Van de Gein, Van der Hoeven, Van der Schoot, Verhelst, Veldhuijzen & Hemker, 2004; Kuhlemeier, Van Til, Hemker, de Klijn, & Feenstra, 2013) showed that two-thirds of the students in grade 6 are not able to compose texts that sufficiently convey a simple message to a reader. The Inspectorate for Education reported that a majority of teachers do not succeed in effectively teaching writing (Henkens, 2010). In the average classroom attention and time devoted to writing are limited: Of the eight hours per week that are spent on language teaching, only 45 minutes are dedicated to writing, most of the time without any form of instruction. Teachers often do not explain how students can approach a writing task, discuss texts, provide feedback, nor do they promote rereading and revising activities (Henkens, 2010). This is partly due to a lack of good materials and tools for teaching, assessing and monitoring writing. The material that is used does not offer teachers enough support to adequately assist their students during the writing process, or to evaluate and assess students' written products (Henkens, 2010). Support for teachers is essential, as during their preservice and in-service professional development they are not sufficiently prepared to teach writing (Pullens, 2012; Van der Leeuw, 2006). In teacher education writing is a neglected topic: During their own professional development teachers hardly received any training in writing or in the teaching of writing (Van der Leeuw, 2007). Thus, to optimize writing education, not only materials for teaching and tools for assessment need to be improved, but also the skills and knowledge of teachers need to be extended. But how can this be accomplished?

BRIDGING THE GAP BETWEEN RESEARCH AND PRACTICE

A considerable amount of research has been conducted aimed to improve writing education, on various aspects. This research has been published in international scientific journals on writing research, such as the Journal of Writing Research, Reading and Writing, Assessing Writing, or in general educational journals, such as the Journal of Educational Psychology or Contemporary Educational Psychology. Moreover, the international book series 'Studies in Writing', recently published volumes on design principles for teaching effective writing and measuring writing. Taken together, all these studies paint a clear picture of the factors that make writing difficult for elementary students and offer possible solutions for the improvement of writing education.

From previous research becomes clear that the first major problem elementary writers face is cognitive overload during writing. During the writing process several resource-demanding cognitive activities have to be performed simultaneously, such as activating prior knowledge, generating content, planning, formulating, and revising, whilst taking into account the communicative goal of the text and the intended audience (Fayol, 1999). Additionally, especially in young writers, the amount of attention required for foundational skills (e.g., handwriting, spelling, and sentence construction) needs to be considered. Developing writers often lack automaticity in these areas (McCutchen, 2011). Due to this limited automaticity, the learner has less attentional capacity for the higher level processes in writing, such as planning, formulating, and revising, which has detrimental effects on text quality (Berninger, Yates, Cartwright, Rutberg, Remy, & Abbott, 1992; McCutchen, 1996).

The second problem elementary writers have to deal with is that, in the way writing education is often organized, learning-to-write and task execution are inextricably linked. For novice writers text production is already so cognitively demanding, that there is hardly any attentional capacity left for learning (Rijlaarsdam & Couzijn, 2000). Thus, to improve students' writing performance, writing instruction should aim to reduce cognitive overload by teaching students skills and knowledge to manage the cognitive activities during writing. Further, writing education can be improved by separating learning from text production. This means that to optimize writing education we should not only address the focus of instruction (what is taught), but also the mode of instruction (how it is taught).

Focus of instruction

The publication of process models of writing in the 1980s (Bereiter & Scardamalia, 1987; Flower & Hayes, 1981) induced a shift from a product-oriented approach to a process-oriented approach of writing. Instead of emphasis on the text and its characteristics, attention should be directed towards the writing process with the underlying assumption that optimizing the process would result in written products of higher quality. This poses a challenge for writing education: What are effective instructional practices for teaching students how to approach a writing task?

Several meta-analyses have been conducted to shed more light on this matter. Some have focused on the effectiveness of a specific type of intervention, such as strategy instruction (Graham, 2006) or the process approach to writing (Graham & Sandmel, 2011). Other meta-analyses have focused on multiple types of interventions with a broad range of students (Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007; Hillocks, 1984), or more specifically on struggling writers (Rogers & Graham, 2008).

However, these meta-analyses do not provide clues on what type of intervention is specifically effective for students in upper elementary grades. These grades are an essential phase in the development of writing skills: the prerequisite foundational skills, such as handwriting, spelling and grammar, have been automatized in such a way that students are able to focus on the composing process itself (Kress, 1994). Moreover, these grades are the stepping stone to secondary education during which writing becomes increasingly important as a tool for learning. Thus, to prepare students for a successful school career it is important that they already master the basics of composition at the end of the elementary school. Therefore it is crucial to examine specifically which instructional practices are effective for this age group.

Mode of instruction

To ameliorate writing education, it is not only important to investigate how the focus of instruction can be improved. It is also necessary to examine how the mode of instruction can be optimized. It does not suffice to provide students with materials and expect them to learn from this. To provide students with optimal learning opportunities, it is important to examine how writing education can be (re)organized in such a way that learning and task execution can be separated. Further, it is important that the instruction is tailored to the student's needs: where one student only needs a little explanation, others may need more support. This support must be offered when needed during all phases of the writing process: in the prewriting stage, during writing, and in the post-writing phase. Adjustment of the mode of instruction is essential to provide students with ample opportunities to learn and practice.

Assessment of writing

Assessing the quality of a text is a tough job. Unlike a sum or multiple-choice question it is not simply a question of right or wrong. Various aspects can contribute to text quality, such as communicative effectiveness, structure, style, spelling, word choice, genre conventions (to name but a few). No wonder that raters can differ enormously in their judgements of what constitutes a good text (cf. Godshalk, Swineford, & Coffman, 1966). To complicate matters even further, the writing performance of student also varies between tasks, depending on genre and/or writing topic (cf. Huang, 2009). This makes it virtually impossible to assess a student's writing proficiency based on one writing task rated by one rater, although this is often the case in classroom practice. There are no standard tests available, forcing teachers to come up with their own solutions. As a result, teachers often do not have a good impression of the writing proficiency of their students, or of the progress students make. To adapt instruction to the students' need and provide effective feedback on students' texts, adequate assessment tools are indispensable. Thus, besides developing teaching material to improve writing education, it is also important develop tools to assess and monitor students' writing.

Professional development of teachers

Any educational reform ultimately relies on teachers (Borko, 2004), which makes teachers the bridge from research to practice. Like with students, it is unrealistic to expect that teachers will improve their teaching automatically when provided with new and improved materials. Therefore, to effectively implement a new teaching program for writing, it is necessary to investigate what additional support should be offered to teachers in order to enlarge their content knowledge about writing as well as pedagogical knowledge on how to teach writing. Further, also should be examined how teachers can effectively apply this knowledge in their daily classroom practice and whether they can transfer their new knowledge and skills to colleagues.

Teacher involvement in research

To induce long-lasting effects in practice, new teaching material should closely match the needs of students as well as teachers. This can be warranted through teacher involvement in all stages of the research process. Teacher involvement in the developmental stages ensures that the material closely matches the interest and developmental stage of the students, as well as the feasibility of the intervention in daily practice. Further, to induce a change in practice, it is essential that teachers deliver the intervention themselves, to their own students.

This poses serious challenges for researchers: to be able to make claims about the effectiveness of an intervention, it is important to establish that the intervention was implemented as intended. Fidelity measures are an essential part of this type of intervention studies, to check if lessons were executed according to plan. Finally it is important to establish the social validity of an intervention, i.e., the acceptability of and satisfaction with the intervention procedures (Luiselli & Reed, 2011). Social validity is a key aspect in the long term effects of interventions: a higher social validity increases the likelihood that (aspects of) an intervention will still be applied after the intervention period has ended. In the case of an intervention aimed to improve classroom practice, teachers might be queried about the complexity of the followed procedure, time involved with the implementation of the intervention, and satisfaction with the outcome (Luiselli & Reed, 2011). This yields important information about the feasibility of the implementation of an intervention in daily classroom practice. Procedures that are perceived as too complicated, impractical or unhelpful, will likely not be adopted.

AIM OF THE DISSERTATION

In this dissertation we aimed to bridge the gap between scientific research and educational practice to improve writing education in the upper grades of elementary school. Building on the findings of previous research, we developed a program for teaching writing, including assessment tools. The effectiveness of the program was tested in two large scale intervention studies in which in total 2785 students and 144 teachers participated, from 125 classes and 52 schools spread over the Netherlands (1420 students and 76 teachers in the first study, and 1365 students and 68 teachers in the second study). We evaluated the effectiveness of the program on the level of the student and the teacher. Regarding the students, we investigated the effect of the intervention as a whole as well as specific components to understand how students' writing can be improved. Regarding the teachers, we examined the impact of the intervention on teachers' self-efficacy, attitudes, and classroom practice. We also investigated the implementation and the social validity of the intervention in order to receive information on the feasibility of implementing the writing program in regular classrooms. and the role of professional development.

CHAPTER OVERVIEW OF THE DISSERTATION

In Chapter 2 to 4 we delineate the framework for our intervention studies. In chapter 2 we first identify effective instructional practices for improving the writing performance by means of a meta-analysis. In this meta-analysis we systematically reviewed 32 (quasi-) experimental writing intervention studies specifically aimed at students grade 4-6 in a general educational setting.

In Chapter 3 we investigated how a student's level of writing proficiency can be determined. We examined how many texts and raters are necessary to make inferences of differences in text quality to differences in writing proficiency. Further, we investigated whether there was an effect of genre on the generalizability of text quality scores of students at the end of primary school. For this, we analyzed text quality scores of texts in different genres, disentangling different variance components, i.e. students, raters, tasks, genre.

To be able to establish students' progress in writing, it is important that his or her performance on a writing task is assessed in a reliable and valid manner. In Chapter 4 we examined the reliability, validity and generalizability of a benchmark rating procedure, by comparing it against a holistic and analytic rating procedure. We also examined whether it is possible to use the same benchmark scale for rating different kinds of texts in a reliable way.

Subsequently, in close collaboration with elementary teachers, we developed Tekster, a comprehensive program for teaching writing. In Tekster we combined several instructional practices into one general overall approach. In Chapter 5 we examined the effectiveness of this program. In a quasi-experimental study, students' writing is measured before and after the intervention program, and compared to a control condition, in order to test whether their writing performance improved over time. To examine the long-term effects of the teaching program, writing performance was also measured two months after following the program.

In Chapter 6, 7, and 8 we take a closer look at the key components of the program, i.e. the writing strategy and feedback. In Chapter 6 we specifically examined the effect of explicitly teaching a writing strategy on students' knowledge of writing, especially on their metacognitive knowledge. Further, we investigated the relationship between knowledge about writing and writing quality in elementary students: did students who possess more knowledge about vbwriting write better texts?

In Chapter 7 we focus on the effectiveness of one important aspect of the writing strategy, i.e., prewriting. In Tekster students were taught to generate ideas and organize these ideas before writing the actual text. We examined the effect of the intervention on students' prewriting behaviour, and the effect of prewriting on students' writing performance, by comparing the effectiveness of different types of prewriting.

Chapter 8 focuses on teachers' feedback practices, as feedback is an important instructional aspect for learning to write. We examined how elementary teachers provide feedback to their students based upon their written texts: did they adjust their feedback to individual students' needs?

In Chapter 9, we examined the added value of a professional development program for the effectiveness of Tekster, the comprehensive program for teaching writing. We also examined whether this professional development program could easily be transferred between colleagues. This was done by examining the differences between teachers who are trained by experts, and teachers who were trained by colleagues in (1) improvement of students' writing performance, (2) perceived self-efficacy for teaching writing, (3) changes in attitudes towards writing and writing instruction, (4) classroom practice, and (5) experiences with the intervention program.

In Chapter 10 we summarize our results, discuss and interpret our findings and the implications of our findings for future research and educational practice.

FINAL REMARK

This dissertation is the result of an intensive collaboration between the two authors. Although we worked closely together at every phase of the research project, we divided some responsibilities along the way according to our expertise. Monica Koster focused mainly on the didactical side of the research project and developed the lessons of the teaching program Tekster together with a team of elementary teachers. Therefore, she is primary responsible for Chapter 2, 6, and 7. Renske Bouwer focused mainly on the assessment of writing, including tools for measuring text quality and providing effective feedback. As a consequence, she is primary responsible for Chapter 3, 4, and 8. We both are equally responsible for the two intervention studies that are described in Chapter 5 and 9.

Due to the fact that the main chapters in this dissertation are set up as separate journal articles some overlap in the method sections and theoretical frameworks is inevitable. This especially applies to the theoretical framework of the two intervention studies in Chapter 5 and 9, and the description of the teaching program Tekster in Chapter 5, 6, 7, and 9. The advantage for the reader is that each chapter can be read on its own.

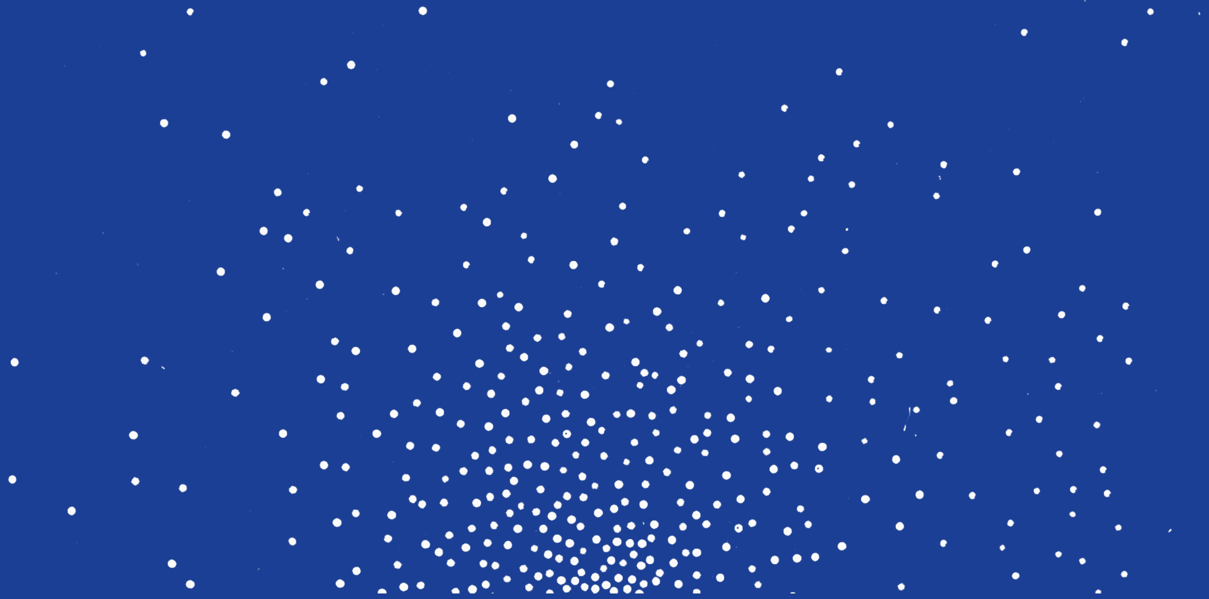
Chapter 2 was published in the Journal of Writing Research and Chapter 3 was published in the journal Language Testing. Chapter 4, 5, 8, and 9 have been submitted for publication and currently have the status 'revise and resubmit'. Chapter 6 and 7 are currently under review for publication.

The results of various chapters have been presented at national and international conferences such as the conferences of EARLI, the European Association for Research on Learning and Instruction (München, 2013; Cyprus, 2015), the EARLI Special Interest Group on Writing (Porto, 2012; Amsterdam, 2014; Liverpool, 2016), Writing Research across Borders (Paris, 2014), and the International Association for the Improvement of Mother Tongue Education (Odense, 2015).

Chapter 2

TEACHING CHILDREN TO WRITE: A META-ANALYSIS OF WRITING INTERVENTION RESEARCH

This chapter has been published as: Koster, M., Tribushinina, E., De Jong, P. F., & Van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research. *Journal of Writing Research, 7*(2), 249-274. doi:10.17239/jowr-2015.07.02.02



It has been established that in the Netherlands, as in other countries, a majority of students do not attain the desired level of writing skills at the end of elementary school. Time devoted to writing is limited, and only a minority of schools succeed in effectively teaching writing. An improvement in the way writing is taught in elementary school is clearly required. In order to identify effective instructional practices we conducted a meta-analysis of writing intervention studies aimed at grade 4 to 6 in a regular school setting. Average effect sizes were calculated for ten intervention categories: strategy instruction, text structure instruction, prewriting activities, peer assistance, grammar instruction, feedback, evaluation, process approach, goal-setting, and revision. Five of these categories yielded statistically significant results. Pairwise comparison of these categories revealed that goal-setting (ES = 2.03) is the most effective intervention to improve students' writing performance, followed by strategy instruction (ES = .96), text structure instruction (ES = .76), peer assistance (ES = .59), and feedback (ES = .88) respectively. Further research is needed to examine how these interventions can be implemented effectively in classrooms to improve elementary students' writing performance.

INTRODUCTION

With the large-scale introduction of computers, tablets, and mobile phones Western society has rapidly become more literate over the last two decades. E-mail and text messages are replacing oral face-to-face and telephone communication, increasing people's need to be able to communicate adequately in writing. Individuals who do not sufficiently master the basic skills of writing will eventually encounter serious problems in participating fully in daily life. More than ever it is essential that children develop their writing competence at a young age, as writing skills play a crucial role in educational and occupational success (National Commission on Writing, 2003).

Despite the fact that composition skills are of vital importance for a successful academic and professional career, it was established that a majority of students in the Netherlands do not attain the desired level of writing skills (Henkens, 2010). A Dutch national assessment study demonstrated that at the end of elementary school (grade 6) most students were not capable of writing texts that sufficiently convey a single, simple message to a reader (Kühlemeier, Van Til, Feenstra, & Hemker, 2013). Further, this study showed that students hardly progress in their writing competencies from grade 4 to grade 6. A national writing assessment in the US yielded similar results: of all grade 8 students only one-third performed at or above proficient level (Salahu-Din, Persky, & Miller, 2008). This is a serious cause for concern, because weaker writers are at a disadvantage in their secondary school and college years, when writing becomes increasingly important as a tool for learning (e.g., Bangert-Drowns, Hurley, & Wilkinson, 2004).

Kühlemeier and colleagues (2013) found that in the Netherlands the time and attention devoted to writing education is limited at elementary school. At the same time the Dutch Inspectorate for Education (Henkens, 2010) concluded that only a minority of schools succeed in effectively teaching writing. Besides, during their own professional education, Dutch teachers do not receive adequate training in writing themselves, nor are they sufficiently prepared for teaching writing (Van der Leeuw, 2006; Smits, 2009). Furthermore, it was established that language teaching materials (i.e., textbooks and teacher manuals) often do not provide sufficient directions for teachers to enable them to support their students' writing processes and to give proper feedback (Stoeldraijer, 2012). It can be concluded that an improvement in the way writing is taught at elementary school in the Netherlands is clearly required.

Above all, any improvement of the teaching of writing in elementary school must be based on interventions that have proven to be effective in enhancing the quality of students' written texts. The aim of this study was to identify effective instructional practices for teaching composition to students in the upper grades of elementary school. An increasing amount of research has been done on writing interventions, resulting in an accumulation of studies testing various instructional approaches. To gain insight into which instructional approaches are specifically effective for elementary students in grade 4 to 6, we conducted a meta-analysis of experimental and quasi-experimental writing intervention studies aimed at students in the upper elementary grades. A meta-analysis is the designated method for this purpose, as the magnitude and the directions of effects of a large number of studies are reviewed in a systematical way.

In the field of writing research, a number of meta-analyses have already been conducted. Some of these analyses focused on a specific type of intervention: for instance, in a review of 39 studies conducted with students from grade 1 to 12, Graham (2006) found that strategy instruction significantly improved students' writing performance. In a meta-analysis on the process approach to writing, Graham and Sandmel (2011) analyzed 29 studies, involving students grade 1 to 12, and found that process writing instruction had a significant, but modest, positive effect on the quality of students' writing. Furthermore, three meta-analyses (Bangert-Drowns et al., 1993; Goldberg, Russell, & Cook, 2003; Morphy & Graham, 2012) investigated the effect of word processing on text quality in grade K to 12, and all found positive effect sizes for this type of treatment, especially for weaker writers.

So far, there have been three comprehensive meta-analyses of experimental and quasi-experimental writing intervention studies, investigating multiple treatments: firstly Hillocks (1984) investigating 60 studies ranging from elementary grades to the first year of college; secondly Graham and Perin (2007) examining 123 writing intervention studies with adolescents (grades 4-12); and, thirdly, Graham, McKeown, Kiuhara, and Harris (2012) analyzed 115 (quasi-) experimental studies involving elementary students, grade 1 to 6. All three analyses used slightly different intervention categories, due to differences in the populations under investigation. Despite this, there was substantial overlap in results. All three meta-analyses consistently found grammar instruction to have a negative effect on text quality with effect sizes [ES] ranging from -.29 (Hillocks, 1984) to -.41 (Graham et al., 2012). Hillocks (1984) and Graham and Perin (2007) both found sentence combining (combining simple sentences into more complex ones), with an ES of .35 and .46 respectively; the study of models (study and imitation of model pieces of writing), with an ES of .22 and .17; and inquiry (present students with data and initiate activities designed to help students develop skills or strategies for dealing with the data in order to write about it), with an ES of .56 and .28, to have a positive effect on students' writing performance. Further, Graham and Perin (2007), as well as Graham and colleagues (2012), found that the process approach to writing, (ES .09 and .40 respectively); strategy instruction, (ES = 1.03 and 1.02); prewriting activities, (ES of .42 and .54); peer assistance when writing, (ES = .70 and .89); setting product goals (ES of 1.00 and .76); and word processing, (ES = .56 and .47), all had a significant positive impact on text quality. In addition, in their elementary meta-analysis, Graham and colleagues (2012) identified seven other effective practices to improve the writing of elementary students writing: feedback (adult and peer), with respective effect sizes of .80 and .37, the use of creativity and imagery (ES = .70), text structure instruction (ES = .59), teaching transcription skills (ES = .55), assessing writing (ES = .42), comprehensive writing programs (ES = .42), and extra writing time (ES = .30).

Rogers and Graham (2008) conducted a meta-analysis of 88 single subject design studies, and found, consistent with the results of the extensive meta-analyses of experimental and quasi-experimental studies, that strategy instruction, word processing, prewriting activities, goal-setting, and sentence construction were effective in improving students' writing performance. Additionally, Rogers and Graham (2008) found that reinforcing writing productivity, teaching strategies for editing text, and teaching strategies for constructing paragraphs were effective for both typical and struggling writers. In contrast to other findings, Rogers and Graham (2008) found a positive effect for the teaching of grammar. As a possible explanation for this divergent finding Rogers and Graham suggested that weaker writers, opposed to typical writers, may have profited from specific grammar instruction, or that the teaching method (teacher modeling) may have contributed positively to the effect of grammar instruction.

The meta-analysis that we conducted can be regarded as a refinement of the previously conducted meta-analyses of writing instruction, as we specifically focused on effective instructional practices for beginning writers (grade 4-6) in a regular educational setting. All previous meta-analyses investigating multiple treatments included a broad range of students: all elementary grades (Graham et al., 2012), adolescents (Graham & Perin, 2007), or elementary to college students (Hillocks, 1984). We expected, however, that different types of treatment would be effective for different groups of students. It was our expectation that the effectiveness of types of intervention would differ between elementary students, secondary students, and college students. Further, we even expected this to differ between lower and upper elementary students. Bourdin and Fayol (1994) have demonstrated that students until the fourth or fifth grade in general perform better orally than in writing, when producing narratives. Their study shows that young students, due to less automation, have to allocate their cognitive resources mostly to the low-level activities of writing, such as lexical access, sentence generating, and graphic execution, which interferes with the higher order skills, such as planning and content generation. Berninger, Yates, Cartwright, Rutberg, Remy, and Abbott (1992) have shown that in the early elementary grades students' writing performance is highly dependent on the degree to which the lower level skills that are conditional to writing are developed. In the upper grades of elementary school it is expected that these lower level skills have been automatized through maturation and practice, in such a way that students are able to focus on the composing process itself (Kress, 1994). It is anticipated that during this stage students will be more sensitive to instruction and practice in basic composition skills. Therefore we decided, unlike Graham and colleagues (2012), to exclude studies aimed at the lower levels of elementary school, and only include studies targeted at students grade 4 to 6.

Further, the prior analyses also included studies targeted at specific groups of students, for example struggling writers, learning disabled students, bilingual students, or high-achieving students. In our opinion, one should be cautious to generalize results from studies targeted at such specific groups to the general population of all students in a regular school setting, as the instructional needs of these groups are bound to differ. For instance, bilingual students might need more grammatical and linguistic support, while the struggling writer might be helped more by instruction in mastering the basics of writing, whereas gifted students could need more challenging writing tasks and approaches. For this reason, we chose to include only studies focused at the full range of students in a regular classroom.

But, above all, none of the previous reviews went beyond summarizing effect sizes and statistically compared interventions to examine whether they differed significantly from each other in effectiveness. In that sense they could be considered statistical reviews more than that they provided answers on the level of differential effectiveness of specific interventions. With our analysis we expanded the previous meta-analyses by not only identifying effective interventions, but by also statistically determining their level of effectiveness by comparison. Lastly, our meta-analysis can also be considered as an update of the previous body of meta-analytical research: a quarter of the studies we located were not included in prior meta-analyses.

In summary, the research question guiding this meta-analysis was: Which instructional practices effectively improve the writing performance of students in the upper elementary grades? To answer this question, we systematically reviewed 32 (quasi-) experimental writing intervention studies aimed at students grade 4-6. The findings from this meta-analysis have important implications for designers of teaching materials and teacher educators, on how the teaching of composition in upper elementary education can be improved.

METHOD

Inclusion criteria and search procedure

In order to be included in the meta-analysis, studies had to meet the following five criteria. First, the study had to involve students in the upper grades of elementary school (grade 4-6) in a regular school setting. Studies that were conducted in a special educational setting or only involving struggling writers were excluded from the analysis. Second, we only included experimental or quasi-experimental studies in which at least two instructional conditions were compared: an experimental condition and a control condition. This could either be a ‘pure’ control condition, in which no extra instruction was given, or a control condition in which an alternative treatment was provided. As a consequence, correlational and qualitative studies were excluded from this meta-analysis. Third, each study had to include a measure of text quality at posttest, as this provided the best indication of the impact of an intervention on students’ writing performance. Scores for text quality are based on a reader’s overall impression of the student’s text, taking into account several factors, such as content, organization, vocabulary, as well as style and tone. Other outcome measures, such as text length or students' motivation were reported only in some of the studies and could therefore not be included in the meta-analysis. Fourth, to be included in the analysis, studies had to provide the statistics necessary to compute a weighted effect size. Lastly, only studies presented in English were included in the meta-analysis.

The studies for this meta-analysis were located by searching the electronic databases of PsychINFO, ERIC and Google Scholar. For our study, we replicated the search procedure employed by Graham and colleagues (2012), using the keywords 'writing' or 'composition', combined with keywords indicating the type of 'intervention', such as: assessment, collaborative learning, creativity, dictation, free writing, genre, goal-setting, grammar, handwriting, imagery, inquiry, mechanics, models, motivation, peer collaboration, peers, planning, prewriting, process approach, process writing, self-evaluation, self-monitoring, sentence combining, sentence construction, spelling, strategies, strategy instruction, summary, technology, word processing, and word processor. Subsequently, we added the following keywords to locate potentially promising practices from recent research: editing, feedback, intervention, modeling, observational learning, outline, outlining, revising, and revision. Further, references of previous meta-analyses, reviews, and obtained papers were examined for relevant studies. Databases of theses, dissertations, and conference proceedings were searched for unpublished studies on the topic. Additionally, a cited reference search of previous reviews and meta-analyses was conducted in Web of Knowledge to identify relevant studies.

This search procedure yielded approximately 2000 results, of which titles and abstracts were closely examined. First, we removed all non-intervention studies, as well as all studies that were not aimed at grades 4-6. Next, we omitted all studies that were not experimental or quasi-experimental. Subsequently, we removed all studies that lacked a proper control condition, and finally we excluded all studies only investigating specific groups of students, such as, for example, struggling writers, learning disabled students, bilingual students, or high-achieving students. We located 37 studies that met all inclusion criteria. However, despite this, five studies did not provide the necessary statistics to calculate effect sizes. We have contacted the authors of these studies to obtain these statistics, but unfortunately received no reply. Regrettably, these studies had to be excluded. The procedure described above resulted in the location of 32 studies that were suitable to be included in our meta-analysis.

Coding procedure

To obtain an adequate description for each study included in the meta-analysis, we coded the following variables: grade, number of participants, description of experimental and control condition, publication type (journal/dissertation/report/conference presentation/paper), and the genre of posttest measure (expository/narrative/informative/persuasive). It should be noted that coding was restricted to posttest measures, as we used these measures to calculate effect sizes.[1] Further, we coded a number of variables of which we expected that they could account for heterogeneity in effect sizes between studies. For this purpose we coded: design of the study (random assignment or quasi-experimental), attrition (% of total sample), period (in days) and intensity (in minutes) of intervention, person providing instruction (researcher, teacher, teaching assistant), and random assignment of teachers to conditions. Due to considerable differences between the scoring procedures that were used, and differences in the interpretation of reliability of scoring, it was not possible to administer one overall reliability score per study. Therefore, we coded aspects of the studies of which is known that they are related to the reliability of writing quality scores: type of assessment of writing quality (holistic or analytical), number of writing tasks at posttest and number of raters assessing the quality of posttest measure (e.g., Rijlaarsdam et al., 2011). All studies were coded by the first author and a trained assistant, a random sample of ten studies (one third of the total sample) were coded by both coders, with 97% agreement.

Categorizing interventions

For the analysis, all studies were thoroughly examined and grouped according to their focus of intervention. Subsequently, studies with a comparable focus of intervention were grouped into categories, based on the categories used in previous meta-analyses (e.g., Graham & Perin, 2007; Graham et al., 2012; Hillocks, 1984). For our study, we maintained the following categories from these meta-analyses: strategy instruction, text structure instruction, peer assistance, process approach, feedback, grammar instruction, and prewriting activities. We decided to use ‘goal-setting’ instead of ‘product goals’, because our sample also included a study involving the setting of process goals, as well as setting product goals. Two types of intervention in our sample did not fit into the categories that were used by previous reviews, therefore we added two new categories: evaluation and revision. This resulted in a total of ten intervention categories, which are summarized in Table 2.1. It should be noted that the intervention categories are not completely mutually exclusive, for instance, prewriting activities and revision also are components of the process approach and strategy instruction. We classified studies according to the main focus of instruction as described by the authors. For example, Bui, Schumaker, and Deshler (2006) characterize their intervention as a strategic writing program, in which they also apply the process approach to writing. As the emphasis of this intervention is on teaching students strategies for writing, it was decided to place this study in the strategy instruction category. Another example of a study of which the intervention has elements of more than one category is the study of Wong, Hoskyn, Jai, Ellis, and Watson

[1] As this information was available for all studies, we restricted ourselves to calculating effect sizes for posttest only. Pretest information was available for 66% of the studies in our sample, whereas 19% of the studies also included a delayed posttest measurement. Pretest measures were only coded to verify whether there were pre-intervention differences between conditions, as most studies had a quasi-experimental design.

Table 2.1 Description of intervention categories

Category	Description
Strategy instruction	Explicit and systematical teaching of writing strategies
Text structure instruction	Explicit teaching of knowledge of the structure of texts
Peer assistance	Students engage in joined activities during (parts of) the writing process
Evaluation	Teaching students to evaluate their own work with specified criteria
Goal-setting	Students are assigned specific product or process goals before writing
Feedback	Students receive comments from others on their writing
Grammar instruction	Explicit teaching of grammar and/or construction of sentences
Revision	Focus on revising draft versions
Prewriting activities	Students engage in activities before writing: generating content/planning
Process approach	Focus on writing process and subprocesses: planning-writing-revising

(2008) which combines self-regulated strategy development with feedback. As the main intervention under investigation is strategy instruction, the study was placed in this category, rather than in the category feedback.

Strategy instruction involves the explicit teaching of strategies for planning, translating and revising. The majority of studies in this category uses the Self Regulatory Strategy Development (SRSD) model of Harris and Graham (1996), in which students are additionally taught self-regulation strategies to manage the writing process, as well as declarative and procedural knowledge about writing. Text structure instruction is the explicit teaching of the structure of a text in a specific genre, such as the organizational structure of a persuasive essay, the story constituents and interrelations of narrative texts, or a compare/contrast essay. Peer assistance involves studies where students have to collaborate during different stages (planning, formulating, or revising) of the writing process, or where some form of tutoring is applied. Evaluation involves teaching students how to reflect on and to assess their own work. Most studies in this category used the 6 (+1) Traits Writing Model, which was developed in the 1980's in the US (Northwest Regional Educational Library, 2013). The 6 (+1) Traits Writing Model asks students to assess their compositions on ideas, organization, voice, word choice, sentence fluency, conventions, and presentation by using reflective questions and rubrics.Goal-setting involves assigning students goals for their writing before they begin: either a product goal (e.g. writing paragraphs), or a process goal (e.g. acquiring a learning strategy). Feedback involves studies in which students receive comments on (aspects of) their writing, either from the teacher or from a peer. Grammar instruction involves interventions that are aimed at the construction of correct sentences. Revision involves studies in which students receive instruction in improving draft versions of texts. Prewriting activities involve studies that focus on techniques for generating content and planning, such as brainstorming, or using graphic organizers. The process approach is a comprehensive intervention where students engage in cycles of planning, formulating, and revising, and in writing for real audiences with real purposes. Instruction is often at individual level, tailored to the student's needs through mini-lessons, writing conferences, and teachable moments. Further, self-reflection and evaluation is stressed, to stimulate student's ownership of their written products. Students collaborate when writing, in a supportive and nonthreatening writing environment (Graham & Sandmel, 2011).

There were three studies in our sample, Arter, Spandle, Culham, and Pollard (1993), Saddler and Graham (2005), and Dejarnette (2008), comparing two intervention conditions. We calculated an effect size for both interventions and subsequently placed them in two intervention categories. Finally, a number of studies investigated multiple conditions, e.g., the study of Schunk and Swartz (1993) investigated the effectiveness of setting product goals, as well as the effectiveness of setting process goals. In these instances we calculated separate effect sizes for all conditions.

Calculation of effect sizes and statistical analysis

For each individual study included in the analysis an effect size was calculated for writing quality at posttest. If a holistic score was available, then this score was used to calculate the effect size. If writing quality was scored on separate aspects, such as organization, ideas, or word choice, separate effect sizes were calculated for each aspect and subsequently averaged into one single effect size. Means and standard deviations were used to obtain effect sizes. Effect sizes were calculated using Hedges' *g* (the standardized mean difference) by subtracting the mean performance of the control group at posttest from the mean performance of the treatment group at posttest, dividing by pooled standard deviation of the two groups. Hedges' *g* provides a slightly better estimate than Cohen's *d*, especially for smaller sample sizes (Borenstein, Hedges, Higgins, & Rothstein, 2011).

For the meta-analysis a random effects model was used, as it was assumed that the true effect varied from study to study, due to differences in participants as well as differences in interventions and implementation of interventions. Rather than estimating one true effect size, a random effects model estimates the mean of a distribution of effects. This allows for generalization to populations beyond the included studies (Borenstein et al., 2011). For each treatment category, an average effect size was calculated as well as the confidence interval and statistical significance of the obtained effect sizes. In this way the effect of various treatments could be compared. Additionally, a test of homogeneity was conducted, to determine whether variability in effect sizes was larger than expected based on sampling error alone. When the homogeneity test was statistically significant, a moderator analysis was conducted to determine whether the variability could be explained by identifiable factors, such as treatment duration, publication type, or grade.

Table 2.2 Description of included studies grouped per intervention category

Study	Publica- tion type	Grade	*N*	Intervention	Genre	Effect size
				Strategy instruction (k=11)		
Brunstein & Glaser (2011)	J	4	115	Strategy instruction + self-regulation vs. strategy instruction	N	0.84
Glaser & Brunstein (2007) 1	J	4	72	Strategy instruction vs. didactic lessons in composition	N	0.48
Glaser & Brunstein (2007) 2	J	4	79	Strategy instruction + self-regulation vs. didactic lessons in composition	N	1.12
Mason et al. (2012) 1	J	4	47	Strategy instruction + self-regulation (TWA + PLANS) vs. no treatment	I	1.13
Bui et al. (2006)	J	5	99	Demand Writing Instruction Model vs. traditional writing instruction (+ Prewriting activities)	n.s.	0.34
Barnes (2013) 1	D	5	178	WISE (Writing In School Every day) vs. no treatment	N,I,P	0.11
Barnes (2013) 2	D	5	189	WISE + professional development vs. no treatment	N,I,P	0.33
Mason et al. (2012)	J	5	48	Strategy instruction (TWA) vs. no treatment	N	0.81
Fidalgo et al. (2013)		6	41	Strategy instruction vs. normal curriculum	I	2.11
Torrance et al. (2007)	J	6	95	CSRI (Cognitive Self Regulation Instruction) vs. normal curriculum	I	3.57
Wong et al. (2008)	J	6	57	SRSD strategy instruction + CHAIR + adult feedback vs. CHAIR + constant training time	P	0.64
				Text structure instruction (k=9)		
Fitzgerald & Teasley (1986)	J	4	49	Instruction in story constituents and interrelations vs. dictionary use and word study	N	1.07
Gordon & Braun (1986)	J	5	54	Instruction in narrative structure vs. instruction in poetry writing	N	0.32
Bean & Steenwyk (1984) 1	J	6	41	Direct instruction rule-governed vs. advice to find main ideas	I	1.07
Bean & Steenwyk (1984) 2	J	6	39	GIST: direct instruction intuitive approach vs. advice to find main ideas	I	0.84
Crowhurst (1990)	J	6	46	Instruction model for persuasion + writing practice vs. group discussion activities	I	1.11
Crowhurst (1991) 1	J	6	50	Instruction model for persuasion + writing practice vs. reading novels and writing book reports	P	1.10
Crowhurst (1991) 2	J	6	50	Instruction model for persuasion + reading practices vs. reading novels and writing book reports	P	0.78
Crowhurst (1991) 3	J	6	50	One lesson persuasion vs. reading novels and writing book reports	P	0.34
Raphael & Kirschner (1985)	C	6	45	Instruction compare-contrast text structure vs. normal curriculum	I	0.26

Study	Publica- tion type	Grade	*N*	Intervention	Genre	Effect size
				Peer assistance (k=9)		
Paquette (2008)	J	4	50	6 + 1 Traits model with cross-age tutoring vs. no extra instruction (+ Evaluation)	n.s.	1.27
Puma et al. (2007) 1	R	4	1249	Writing Wings (cooperative writing) vs. normal curriculum	N,I	0.07
Saddler & Graham (2005) 1	J	4	44	Sentence combining with peer assistance vs. grammar instruction	N	1.66
Puma et al. (2007) 2	R	5	347	Writing Wings (cooperative writing) vs. normal curriculum	N,I	0.03
Yarrow & Topping (2001) 1	J	5	14	Metacognitive strategy instruction with peer assistance (tutor) vs. metacognitive strategy instruction with no interaction	N	0.70
Yarrow & Topping (2001) 2	J	5	12	Metacognitive strategy instruction with peer assistance (tutee) vs. metacognitive strategy instruction with no interaction	N	0.52
Brakel Olson (1990) 2	J	6	41	Writing lessons + peer partner vs. writing lessons only	N	0.42
Hoogeveen (2013) 1	D	6	96	Specific genre knowledge + peer response vs. no extra instruction	N,E	1.11
Hoogeveen (2013) 2	D	6	93	General aspects of communicative writing + peer response vs. no extra instruction	N,E	0.30
				Evaluation (k=7)		
Collopy (2009)	J	4	100	6 Traits writing model vs. no extra instruction	N	0.31
Paquette (2008)	J	4	50	6 + 1 Traits model with cross-age tutoring vs. no extra instruction (+ Peer assistance)	n.s.	1.27
Tienken & Achilles (2003)	J	4	98	Skills and strategies to self-assess writing vs. no extra instruction	N	0.41
Ross et al. (1999)	J	4/5/6	296	Self-evaluation with rubrics + teacher feedback vs. normal curriculum development	N	0.74
Arter et al. (1994) 1	C	5	132	6 Traits writing model vs. observation (normal curriculum) (+ Process approach)	E,N	0.20
DeJarnette (2008)	D	5	131	6 + 1 Traits writing model vs. Writing workshop	N	0.73
Coe et al. (2011)	R	5	4134	6 Traits writing model vs. no extra instruction	E	0.01
				Goal-setting (k=6)		
Schunk & Swartz (1993) 2	J	4	20	Process goal + progress feedback vs. general goal (+ Feedback)	E,N,I	3.03
Schunk & Swartz (1993) 2	J	4	20	Process goal vs. general goal	E,N,I	2.62
Schunk & Swartz (1993) 2	J	4	20	Product goal vs. general goal	E,N,I	1.05
Schunk & Swartz (1993) 1	J	5	30	Process goal + progress feedback vs. general goal (+ Feedback)	E,N,I	3.15

Study	Publica- tion type	Grade	*N*	Intervention	Genre	Effect size
Schunk & Swartz (1993) 1	J	5	30	Process goal vs. general goal	E,N,I	2.66
Schunk & Swartz (1993) 1	J	5	30	Product goal vs. general goal	E,N,I	1.65
				Feedback (k=4)		
Schunk & Swartz (1993) 2	J	4	20	Process goal + progress feedback vs. general goal (+ Goal-setting)	E,N,I	3.03
Schunk & Swartz (1993) 1	J	5	30	Process goal + progress feedback vs. general goal (+ Goal-setting)	E,N,I	3.15
Holliway (2004) 1	J	5	55	Feedback + rating vs. one sentence feedback	E	0.84
Holliway (2004) 1	J	5	48	Feedback + reading as the reader vs. one sentence feedback	E	0.69
				Grammar instruction (k=4)		
Saddler & Graham (2005) 1	J	4	44	Grammar instruction vs. sentence combining with peer assistance	N	-1.66
Gein (1991) 1	D	4	109	School grammar vs. direct writing	E,N	-0.05
Gein (1991) 2	D	4	110	Sentence construction vs. direct writing	E,N	0.06
Gein (1991) 3	D	4	111	School grammar vs. sentence construction	E,N	-0.11
				Revision (k=3)		
Brakel Olson (1990) 1	J	6	40	Revision instruction vs. no extra instruction	N	0.04
Brakel Olson (1990) 3	J	6	37	Revision instruction + peer partner vs. no extra instruction (+ Peer assistance)	N	0.85
Fitzgerald & Markham (1987)	J	6	30	Revision instruction vs. reading good literature	N	0.89
				Prewriting activities (k=3)		
Brodney et al. (1999) 1	J	5	51	Reading combined with prewriting vs. no extra instruction	E	0.93
Brodney et al. (1999) 3	J	5	49	Prewriting only vs. no extra instruction	E	0.17
Bui et al. (2006)	J	5	99	Demand Writing Instruction Model vs. traditional instruction (+ Strategy instruction)	n.s.	0.34
				Process approach (k=3)		
Arter et al. (1994)	C	5	132	Process approach vs. 6 Traits model (+ Evaluation)	E,N	-0.20
DeJarnette (2008) 2	D	5	131	Writing workshop vs. 6 + 1 Traits writing model (+ Evaluation)	N	-0.73
Varble (1990)	J	6	128	Whole language group vs. traditional language instruction	I	0.16

Note. For Study, numbers behind the references indicate that effect sizes were calculated for multiple conditions, or groups; these effect sizes are reported separately. For Publication type, J: Journal, D: Dissertation, R: Report, C: Conference presentation, P: Paper. For Genre: E: Expository, N: Narrative, I: Informative.

Description of studies included in the meta-analysis
Table 2.2 contains a description of all studies included in the analysis and their effect sizes, grouped per intervention category. The intervention categories are ranked according to the amount of effect sizes they contain, starting with strategy instruction as the largest category (11 effect sizes). Within the categories, studies are arranged per grade, in alphabetical order.

For each study the following information is given: reference, publication type, grade, number of participants, short description of intervention and control condition, genre of text written at posttest measure, and the effect size. As can be seen, there are seven categories containing four or less effect sizes. We acknowledge the fact that these sample sizes do not allow for firm conclusions. Nevertheless, for the sake of completeness, it was decided to retain these categories in the analysis, as this would at least provide an indication of the possible efficacy of these types of interventions. In total, we calculated 55 effect sizes from 32 studies, and divided them into 10 intervention categories.

RESULTS

First, a random effects model was used to obtain an overall average effect size for all studies included in the meta-analysis. This overall effect size was $g = .72$, with a 95% confidence interval of [.49 - .94]. As effect sizes are highly dependent on study characteristics, additional analysis was needed to establish whether the various effect sizes together in the sample provide a proper estimate of the effect size in the population. This can be determined by conducting a homogeneity test. This test indicates if the variability in effect sizes is larger than the expected variability based on sampling error alone. As the studies in our sample varied widely in focus and approach, we expected significant heterogeneity, which was confirmed by the homogeneity test: $Q = 511.51$, $df = 54$, $p < .001$. This indicated that a common effect size for the total sample of studies could not be assumed.

First, we investigated possible publication bias by conducting a moderator analysis with publication type as moderator on all studies in the meta-analysis. This analysis yielded no significant result ($p = .22$), indicating that the effect sizes of studies published in peer reviewed journals did not differ systematically in their effect sizes from studies from other publication types. The next step in our analysis was to examine the effectiveness of the various intervention categories, by including these 10 categories in our model as explanatory variables. The inclusion of the intervention categories significantly improved the model, according to a likelihood ratio test, with $X^2 = 19.69$, $df = 9$, $p < .001$. This means that differences in effect sizes were (at least partly) explained by the type of intervention.

Table 2.3 gives the summary statistics for all intervention categories, presented in the same order as Table 2.2. These statistics include, per intervention category, the number of effect sizes, the average effect size and standard error, the 95% confidence interval, and the heterogeneity statistics Q (test statistic for heterogeneity) and I^2 (percentage of total heterogeneity/variability). As can be seen in Table 2.3, we found two negative effects, for grammar instruction and the process approach. These interventions did not improve the quality of students' writing. However, all other main effects were positive. Of these positive effects, five main effects significantly deviated from zero. These were, in order of effect size: goal-setting, strategy instruction, feedback, text structure instruction, and peer assistance. Post-hoc analysis was conducted by a contrast analysis in which all interventions were compared pairwise. Results from these analyses showed that goal-setting was by far the most effective intervention ($X^2 \geq 36.81$, $df = 1$, $p < .001$). However, as can be seen in Table 2.2, all effect sizes in the category goal-setting were calculated from one study in which multiple conditions and grades were compared (Schunk & Swartz, 1993). This result should therefore be interpreted with caution. Goal-setting was followed by strategy instruction ($X^2 \geq 26.06$, $df = 1$, $p < .001$), text structure instruction ($X^2 \geq 12.82$, $df = 1$, $p <$ $.001$), and peer assistance ($X^2 \geq 7.64$, $df = 1$, $p = .006$) respectively. These three categories were all based on nine or more effect sizes from different studies. Feedback also proved to be an effective intervention, however, not more effective than prewriting activities.

Table 2.3 Summary of statistics for intervention categories

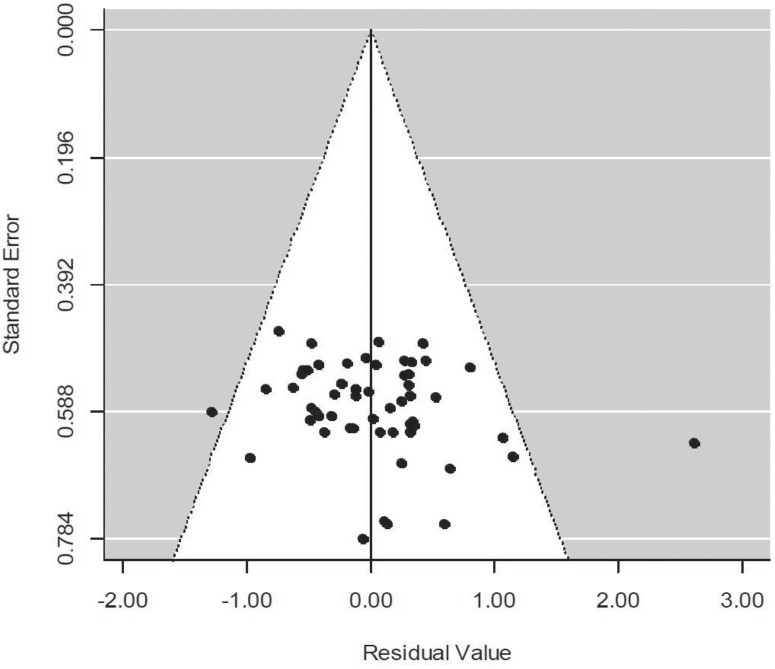
Intervention category	*N*	Average	*SE*	95% Confidence interval		Heterogeneity	
				Lower	Upper	Q	I^2
Strategy instruction	1	0.96 ***	0.19	0.59	1.33	109.99 ***	94.30
Text structure instruction	9	0.76 ***	0.21	0.34	1.18	11.91	33.87
Peer assistance	9	0.59 **	0.21	0.17	1.01	56.05 ***	89.83
Evaluation	7	0.43	0.23	-0.01	0.87	66.56 ***	87.57
Goal-setting	6	2.03 ***	0.33	1.37	2.68	13.47 *	62.61
Feedback	4	0.88 *	0.38	0.14	1.61	25.08 ***	91.08
Grammar instruction	4	-0.37	0.30	-0.97	0.22	20.16 ***	91.84
Revision	3	0.58	0.38	-0.17	1.33	4.14	51.59
Prewriting activities	3	0.13	0.36	-0.58	0.85	3.91	48.57
Process approach	3	-0.25	0.34	-0.92	0.41	12.78 **	84.58

Note. *** p < .001, ** p < .01, * p < .05

The homogeneity test indicated that there was still a significant amount of residual heterogeneity in the sample ($QE = 283.18$, $df = 45$, $p < .001$). Therefore, we inspected the funnel plot (see Figure 2.1) to locate outliers that could be a potential source of heterogeneity. A funnel plot is a scatterplot of the intervention effect against a measure of study size. In the funnel plot in Figure 2.1 the residuals of the model with the intervention categories as explanatory variables were plotted against the standard error. The straight lines in Figure 2.1 define the region in which 95% of the studies was expected, in the absence of homogeneity. It can be seen that the studies were more or less symmetrically spread around the overall average effect size, and that most points were located in the region between the straight lines. This was an indication that there was no systematic heterogeneity in our sample. Two outliers (6.25% of the total sample) were located. The forest plot that we subsequently created (see Appendix A), identified these outliers to be the studies of Torrance et al. (2007), and Saddler and Graham (2005). The effect size in the study of Torrance et al. (2007) was underestimated in the analysis whereas the observed effect size in the study of Saddler and Graham (2005) was smaller than expected, which meant that the effect size of the first study was larger, whereas in the latter study the effect size was smaller than in comparable studies (see also Figure 2.1). The analysis was repeated without these studies, but the outcome of this analysis did not significantly differ from the previous analysis ($X^2 = 3.61$, $df = 2$, $p = 0.16$). Hence, it was decided to maintain these studies, and to retain the previously estimated model for further analysis.

Subsequently, a moderator analysis was conducted to examine whether the variability between studies could be attributed to one or more identifiable factors. We examined whether there were systematical differences in effect sizes between studies with a proper control condition and studies comparing different intervention conditions. In six intervention categories there were one or more

Figure 2.1 Funnel plot of final model.



studies without a no extra instruction control condition. Contrary to expectations, the inclusion of control condition as a moderating variable did not result in a significant reduction of residual heterogeneity ($QE = 220.37$, $df = 37$, $p < .001$), and for none of the intervention categories the parameter estimate for control condition was significant (p-values ranging from .29 to .90). Next, grade, duration of intervention, type of assessment of writing quality (holistic or analytical), number of writing tasks in posttest, and number of raters assessing the quality of posttest measure were considered as moderating factors. None of these factors significantly reduced the heterogeneity between studies in the total sample.

In the next step of the analysis, closer examination of the intervention categories separately revealed no significant heterogeneity in four categories: text structure instruction, process approach, revision, and prewriting (p-values ranging from .08 to .16). We further investigated the heterogeneity within the remaining intervention categories. This analysis was limited to categories containing more than five effect sizes, i.e. strategy instruction, peer assistance, evaluation and goal-setting, as in the smaller categories the heterogeneity can largely be attributed to differences between individual studies. In the larger categories systematic factors may have caused heterogeneity, and this was examined by performing a moderator analysis on the separate categories with grade, duration of intervention, type of assessment of writing quality (holistic or analytical), number of writing tasks in posttest, and number of raters assessing the quality of posttest measure as potential moderators.

In strategy instruction, grade appeared to be a significant moderator: effect sizes were systematically larger in grade 6 (2.19) than in either grade 4 or 5 (0.59). Further, we found that effect sizes in this category were smaller (-0.86) for studies in which text quality at posttest was assessed analytically compared to studies in which holistic assessment was applied. In the category evaluation, genre of posttest was a significant moderator: effect sizes were smaller (-0.11) for expository texts. In the category peer assistance, heterogeneity could largely be attributed to one large study (Puma et al., 2007) with a relatively low effect size. In goal-setting, heterogeneity could be attributed to differences between conditions.

However, from the 95% confidence interval statistics reported in Table 2.3 can be concluded that, despite significant heterogeneity within the categories of interventions that significantly improve writing proficiency, the effects in these categories were still largely positive, even at the lower bound of the confidence interval.

DISCUSSION

Effective interventions to improve elementary students' writing

It has been established that, in the Netherlands, the way writing is taught in elementary school needs to be improved. The aim of this meta-analysis was to identify evidence-based effective instructional practices for teaching writing to students in grade 4 to 6. To determine this, we calculated average effect sizes for 10 types of interventions. The results show that the most effective interventions to improve students' writing are, in order of effect sizes: goal-setting, strategy instruction, text structure instruction, feedback, and peer assistance. Post hoc analysis demonstrates that goal-setting is the most effective intervention, followed by strategy instruction, text structure instruction, peer assistance, and feedback. This is in line with the findings of recent previous reviews (Graham & Perin, 2007; Graham et al., 2012), even though we limited our analysis to students in grade 4 to 6 in a regular educational setting. However, our findings are corroborated by statistical analysis.

The results of our analysis show that goal-setting was by far the most effective intervention. However, as stated before, it should be noted that all effect sizes in this category come from one (twenty year old) study (Schunk & Swartz, 1993), comparing multiple conditions and multiple grades. Thus, these results only allow for tentative conclusions. Support for the positive effect of setting product goals can be found in previous meta-analyses (Graham & Perin, 2007; Graham et al., 2012) albeit for (partly) different populations of students (special needs learners, struggling writers, and slightly older students). This indicates that setting goals could help to improve students' writing.

Strategy instruction is the next effective intervention. Strategy instruction is the largest intervention category in our analysis, which allows for robust conclusions. Of all types of intervention, strategy instruction is by far the most investigated. It should be noted that that the majority of studies in this category examined the self-regulated strategy development (SRSD) approach of Harris and Graham (1996) to strategy instruction or a variation thereof. The SRSD approach seems to have developed into the 'standard' in strategy instruction, which is hardly surprising as studies examining SRSD invariably yield large effect sizes. Previous meta-analyses (Graham, 2006; Graham & Perin, 2007; Graham et al., 2012) also found SRSD to be a highly effective intervention for all types of learners (struggling writers, learning disabled, average, gifted) in a wide range of grades (grade 2 to 10). A subsequent moderator analysis, which we performed in all categories containing more than five effect sizes from different studies, shows that, in our sample, in grade 6 the (average) effect of strategy instruction appears to be much higher than in either grade 4 or 5. A possible explanation for this finding may be that in grade 6 students' lower level skills have been developed to such an extent that they profit the most from the explicit teaching of writing strategies. Further, we find that effect sizes in this category are smaller in studies where text quality is assessed analytically, compared to studies in which holistic assessment is used. In analytical assessment scoring rubrics are used: a set of criteria and standards that are linked to the learning objectives of the task at hand. Therefore, analytical assessments are more task-specific than a holistic assessment, which makes them harder to generalize to writing proficiency (Schoonen, 2005; Rijlaarsdam et al., 2011). As all different aspects of a text are evaluated separately, and subsequently combined into one final total score, analytical scores tend to be lower than holistic scores (Schoonen, 2005).

The next effective intervention category is text structure instruction. This category is a homogeneous sample of studies. The studies in this category investigate the effect of explicit teaching of (elements of) text structure, in different types of texts: narrative, persuasive, and compare-contrast texts. In all studies in this category the explicit teaching of text structure leads to a significant improvement of students' writing performance.

Text structure instruction is followed by peer assistance. Peer assistance is a diverse category: collaboration between students is applied in different phases of the writing process, with diverse types of interventions. As can be seen in Table 2.2, the effect of peer assistance depends on how it is applied, and on the focus of the intervention. Studies with mainly cooperative writing (e.g. Puma et al., 2007) have smaller effects than studies combining peer assistance with more targeted types of interventions, such as the teaching of specific genre knowledge (Hoogeveen, 2013) or sentence combining (Saddler & Graham, 2005). Peer tutoring is also an effective practice to improve students' writing, as is shown by the study of Yarrow and Topping (2001). This study further shows that the writing scores of tutors improved more than those of the tutees. An explanation for this result may be that students learn more from explaining the material to others: you can only adequately explain something if you understand it yourself.

With only four effect sizes from two studies, feedback is one of the smaller intervention categories. Although seemingly effective, more research is needed to allow for more robust conclusions, as feedback can take many forms (e.g. peer feedback vs. teacher feedback) and can be applied in different ways (e.g. product-focused vs. process-focused). Further research should examine how and in what form feedback can be applied in teaching writing to improve students' writing performance.

Grammar instruction and the process approach to writing yield negative average effect sizes. The negative effect for grammar instruction confirms the findings in previous meta-analyses (Graham & Perin, 2007; Graham et al., 2012; Hillocks, 1984). Apparently, attention for the construction of correct sentences does not lead to improvement in text quality. This may be due to lack of transfer effects: when grammar is taught in isolation, and not in a 'real' writing context, it may not be clear to students how to apply what they learned when writing a text.

The negative effect for process approach may be explained by several factors. First, it is a small, but nevertheless homogeneous, intervention category of only three studies. In two out of these three studies, process approach is the control condition, thus compared with another (in this case: more effective) intervention type. We suspected that this could have resulted in lower effect sizes than when process approach would have been compared to a 'pure' control group. However, our suspicions were not confirmed by subsequent analysis with type of control condition as a moderator. There are several possible explanations for this result: the most straightforward one is that there are indeed no differences, but it can also be that our sample is too small and therefore lacking power to reveal systematic differences. However, it can also be that the process approach is too comprehensive for beginning writers: working on too many aspects at the same time. Beginning writers may profit more from a targeted intervention, such as text structure or strategy instruction. It must be noted that Graham and Perin (2007) found a (small) positive effect, for the process approach in their meta-analysis for adolescent students. This might indicate that the process approach is an effective approach for teaching writing to more experienced writers, but that this approach is less suitable for beginning writers.

Limitations of the study

We recognize the fact that some categories were small (≤ 4 effect sizes) and therefore only allow for tentative conclusions on the overall effectiveness of these intervention categories. Nevertheless, since we wanted to obtain as much information as possible from the available data, these categories were included in the analysis, to examine their potential effectiveness.

A complicating factor in interpreting the results of the analysis is the fact that there was a considerable heterogeneity between studies that could not fully be accounted for by identifiable factors. However, it should be noted that the heterogeneity is overestimated due to the amount of small studies in our sample. A large amount of small studies in a category results in considerably more heterogeneity between studies, whereas in larger studies there is more heterogeneity within the study, and less between studies. In our sample, we see that the heterogeneity in the smaller categories is often caused by differences between individual studies. For instance, variation between studies may have been caused by differences in operationalization, such as the materials that were used, and the instruction that was given. Further, the number and nature of assignments that students had to work on varied considerably: from one writing task in one genre, to several writing tasks in various genres. The period of intervention varied even more: from one day to one year. A complicating factor in the analysis is that key aspects, such as the control condition, the nature of the posttest, the exact period of intervention, and the exact time spent on the intervention were not always described explicitly, which made it troublesome to code for these variables. These aspects can contribute to heterogeneity, but they cannot be included in a meta-analysis in a meaningful way if they are not reported accurately.

Suggestions for further research

From our study it becomes clear that there is not much writing intervention research conducted for students in the upper grades of elementary school. We can conclude that more research in this area is clearly needed. Some of the intervention categories in our meta-analysis were too small to draw firm conclusions about their effectiveness. Especially in these categories more research should be conducted. Particularly the effectiveness of goal-setting needs further investigation, as our results indicate that it might be very effective in improving writing. It would certainly be worthwhile to investigate whether the positive results of the study of Schunk and Swartz (1993) can be replicated in other studies. But also feedback and prewriting activities need to be studied more closely. Additionally it should be examined if a combination of highly effective interventions will lead to even better student outcomes. In other words: is it worthwhile to add one highly effective intervention to another highly effective intervention, or does this only lead to marginal improvement? Further, other types of interventions, and new approaches should be developed and tested.

Furthermore, of the studies in our sample, 34% employs a posttest-only design and 47% a pretest-posttest design in which the effect is measured directly at the end of an intervention. However, to make substantiated claims about the effectiveness of an intervention, a delayed posttest should be included to measure retention. Often, the posttest closely resembles what is taught during the intervention, which may lead to an overestimation of the effects. A delayed posttest could provide more information on the long-term effects of interventions on students' writing. Therefore, to make any claims about the 'real' effectiveness of interventions, delayed posttest data are essential. Unfortunately, this is still not common practice in intervention research.

Recommendations for teaching

This meta-analysis provides some valuable clues as to what works in teaching writing. Certainly more in-depth research is needed into what specifically works and what not, but we were already able to identify promising interventions for successfully teaching writing to students in the upper grades of elementary education. On the basis of our results, we must conclude that, to successfully improve the quality of writing of beginning writers, the writing curriculum should include goal-setting, strategy instruction, text structure instruction, feedback and peer interaction. Setting process goals, such as learning to apply a certain strategy, was highly effective. Strategy instruction was more effective when combined with teaching self-regulatory skills. Overall, we found that specific, targeted interventions, such as explicit instruction in applying strategies or how to structure a text were particularly effective for elementary students. What we still do not know, is what the ideal instructional program for teaching composition skills should look like: which materials should we use, how much students have to write, how much practice students need, how we support the students' writing process, how we give appropriate feedback, and so on. In that respect, this analysis provides only rough guidelines for teaching, not a ready to use panacea. To determine what really works, extensive testing in classrooms is still needed.

REFERENCES OF STUDIES INCLUDED IN THE META-ANALYSIS

Arter, J. A., Spandel, V., Culham, R., & Pollard, J. (1994). *The impact of training students to be self-assessors of writing.* New Orleans. Paper presented at AERA.

Barnes, J. C. (2013). *The effects of a writing intervention on fifth-grade student achievement* (Doctoral dissertation).

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Literacy Research, 16*(4), 297-306.

Brakel Olson, V.L. (1990). The revising processes of sixth-grade writers with and without peer feedback. *The Journal of Educational Research, 84*(1), 22-29.

Brodney, B., Reeves, C., & Kazelskis, R. (1999). Selected prewriting treatments: Effects on expository compositions written by fifth-grade students. *The Journal of Experimental Education, 68*(1), 5-20.

Brunstein, J. C., & Glaser, C. (2011). Testing a path-analytic mediation model of how self-regulated writing strategies improve fourth graders' composition skills: A randomized controlled trial. *Journal of Educational Psychology, 103*(4), 922-938.

Bui, Y. N., Schumaker, J. B., & Deshler, D. D. (2006). The Effects of a Strategic Writing Program for Students with and without Learning Disabilities in Inclusive Fifth- Grade Classes. *Learning Disabilities Research & Practice, 21*(4), 244-260.

Coe, M., Hanita, M., Nishioka, V., & Smiley, R. (2011). *An Investigation of the Impact of the 6+ 1 Trait Writing Model on Grade 5 Student Writing Achievement. Final Report. NCEE 2012-4010.* National Center for Education Evaluation and Regional Assistance.

Collopy, R. M. (2008). Professional development and student growth in writing. *Journal of Research in Childhood Education, 23*(2), 163-178.

Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education/Revue canadienne de l'éducation, 15*(2), 155-172.

Crowhurst, M. (1991). Interrelationships between reading and writing persuasive discourse. *Research in the Teaching of English*, 314-338.

Danoff, B., Harris, K. R., & Graham, S. (1993). Incorporating strategy instruction within the writing process in the regular classroom: Effects on the writing of students with and without learning disabilities. *Journal of Literacy Research, 25*(3), 295-322.

DeJarnette, N. K. (2008). *Effect of the 6+ 1 Trait Writing Model on Student Writing Achievement.* ProQuest.

Fidalgo, R., Torrance, M., Rijlaarsdam, G. & Van den Bergh, H. (2013). *Social learning in strategy-focused writing instruction: A components analysis.* Manuscript submitted for publication.

Fitzgerald, J., & Teasley, A. B. (1986). Effects of instruction in narrative structure on children's writing. *Journal of Educational Psychology, 78*(6), 424-432.

Fitzgerald, J., & Markham, L. R. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4*(1), 3-24.

Glaser, C., & Brunstein, J. C. (2007). Improving fourth-grade students' composition skills: Effects of strategy instruction and self-regulation procedures. *Journal of Educational Psychology, 99*(2), 297-310.

Gordon, C. J., & Braun, C. (1986). Mental processes in reading and writing: A critical look at self-reports as supportive data. *The Journal of Educational Research,* 292-301.

Holliway, D. R. (2004). Through the eyes of my reader: A strategy for improving audience perspective in children's descriptive writing. *Journal of Research in Childhood Education, 18*(4), 334-349.

Hoogeveen, M. (2013). *Writing with peer response using genre knowledge* (Unpublished doctoral dissertation). SLO: Enschede.

Knudson, R. E. (1991). Effects of instructional strategies, grade, and sex on students' persuasive writing. *The Journal of Experimental Educational,* 141-152.

Mason, L. H., Davison, M. D., Hammer, C. S., Miller, C. A., & Glutting, J. J. (2012). Knowledge, writing, and language outcomes for a reading comprehension and writing intervention. *Reading and Writing,* 1-26.

Paquette, K. R. (2008). Integrating the 6+ 1 writing traits model with cross-age tutoring: An investigation of elementary students' writing development. *Literacy Research and Instruction, 48*(1), 28-38.

Puma, M., Tarkow, A., & Puma, A. (2007). *The challenge of improving children's writing ability: A randomized evaluation of "Writing Wings".* Institute of Education Sciences. Retrieved September 2, 2013 from: http://eric.ed.gov/?id=ED504279.

Raphael, T. E., & Kirschner, B. M. (1985). *The effects of instruction in compare/contrast text structure on sixth-grade students' reading comprehension and writing products.* Research Series No. 161.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6*(1), 107-132.

Saddler, B., & Graham, S. (2005). The effects of peer-assisted sentence-combining instruction on the writing performance of more and less skilled young writers. *Journal of Educational Psychology, 97*(1), 43.

Schunk, D. H., & Swartz, C. W. (1993). Goals and progress feedback: Effects on self-efficacy and writing achievement. *Contemporary Educational Psychology, 18*(3), 337-354.

Tienken, C. H., & Achilles, C. M. (2003). Changing teacher behavior and improving student writing achievement. *Planning and Changing,* 34(3), 153-168.

Torrance, M., Fidalgo, R., & García, J. N. (2007). The teachability and effectiveness of cognitive self-regulation in sixth-grade writers. *Learning and Instruction, 17*(3), 265-285.

Varble, M. E. (1990). Analysis of writing samples of students taught by teachers using whole language and traditional approaches. *The Journal of Educational Research,* 245-251.

Wong, B. Y., Hoskyn, M., Jai, D., Ellis, P., & Watson, K. (2008). The comparative efficacy of two approaches to teaching sixth graders opinion essay writing. *Contemporary Educational Psychology, 33*(4), 757-784.

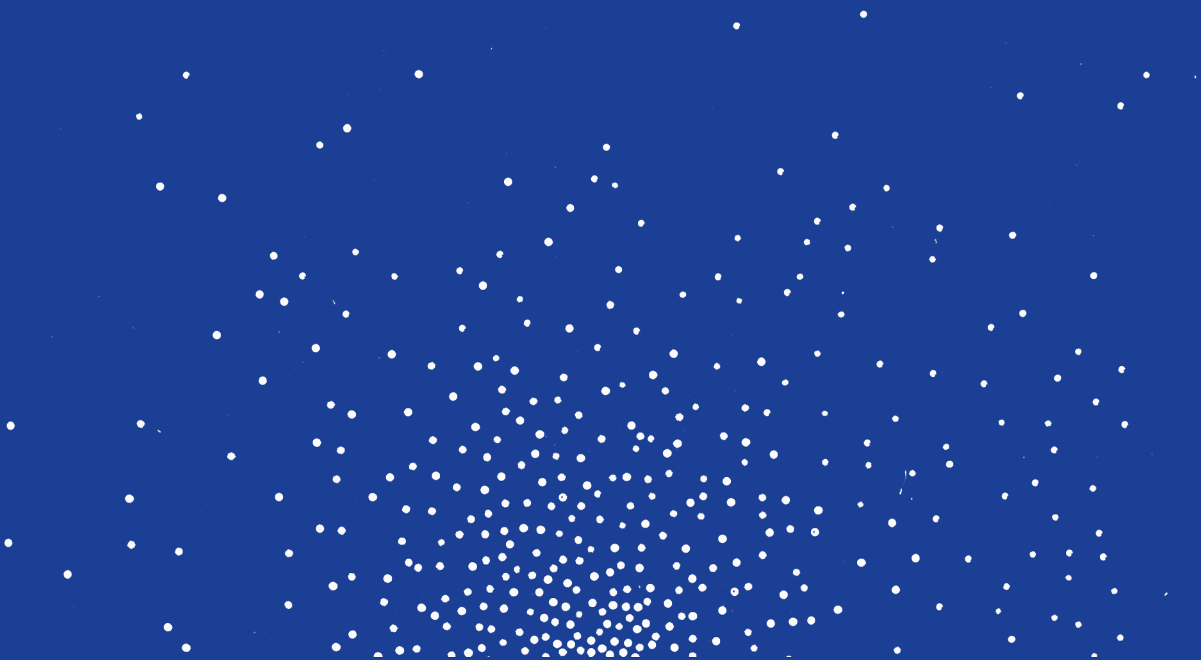
Yarrow, F., & Topping, K. J. (2001). Collaborative writing: The effects of metacognitive prompting and structured peer interaction. *British Journal of Educational Psychology, 71*(2), 261-282.

Chapter 3

EFFECT OF GENRE ON THE GENERALIZABILITY OF WRITING SCORES

This chapter has been published as: Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2016). Effect of genre on the generalizability of writing scores. *Language Testing, 32*(1), 83-100. doi:1010.1177/0265532214542994

An adapted version of this chapter, translated in Dutch, has been published as: Bouwer, R., & Van den Bergh, H. (2016). Toetsen van schrijfvaardigheid: Hoeveel beoordelaars, hoeveel taken? *Levende Talen Tijdschrift, 16* (3), 3-12.



In the present study, aspects of the measurement of writing are disentangled in order to investigate the validity of inferences made on the basis of writing performance and to describe implications for the assessment of writing. To include genre as a facet in the measurement, we obtained writing scores of 12 texts in four different genres for each participating student. Results indicate that across raters, tasks and genres, only 10% of the variance in writing scores is related to individual writing skill. In order to draw conclusions about writing proficiency, students should therefore write at least three different texts in each of four genres rated by at least two raters. Moreover, when writing scores are obtained through highly similar tasks, generalization across genres is not warranted. Inferences based on text quality scores should, in this case, be limited to genre-specific writing. These findings replicate the large task variance in writing assessment as consistently found in earlier research and emphasize the effect of genre on the generalizability of writing scores. This research has important implications for writing research and writing education, in which writing proficiency is quite often assessed by only one task rated by one rater.

INTRODUCTION

Assessment of writing proficiency is essential to writing education as well as writing research. For example, teachers want to know whether their students are able to write well-structured and understandable texts and researchers want to know whether their writing intervention was effective. As is the case for all performance assessments, the most appropriate way to assess writing proficiency is to have people write one or more texts (Huot, 1990b). It is, however, hard to generalize text quality scores to writing skills in general, as ratings of text quality do not only vary due to individuals' writing skills, but also due to characteristics of the measurement situation such as raters and tasks.

The effect of raters on text quality scores has been studied extensively in earlier research, but the effects of task are less understood. Analyses of task effects have almost always been based on multiple tasks in one genre, or on single tasks within multiple genres, such as narrative and argumentative writing. Hence, topic and genre effects are confounded. It is still unclear whether generalization of writing scores is warranted across genres when writing assessments only include highly similar tasks (e.g., Van den Bergh, Maeyer, Van Weijen, & Tillema, 2012). On the other hand, effects of genre cannot be inferred when tasks differ both in genre and topic (Coffman, 1966; Veal & Tillman, 1971). In the current study, the effects of the writing task are further disentangled, allowing for a more valid interpretation of the results of writing assessments. The dual aim of this study is to (a) investigate and demonstrate the validity of inferences made on the basis of writing performance both within and across genres and (b) describe its implications for the assessment of writing proficiency.

One of the facets in the measurement that causes variance in text quality ratings, other than individuals' writing skills, is the rater. Raters are not always consistent in their judgments and they often disagree (Godshalk, Coffman, & Swineford, 1966; Schoonen, Vergeer, & Eiting, 1997). Rater variability may impact both absolute decisions (i.e., decisions concerning performance levels) and relative decisions (i.e., decisions concerning the ranking of students) that are made on the basis of writing performance. For instance, some raters are more strict than others (Weigle, 2002). Ratings of strict raters are consistently too harsh, in comparison to other raters or established benchmarks. When rater severity is not taken into account, student's writing performance will, in this case, be underestimated. Score variance may also be affected by interactions between rater and student. For instance, raters who differ in their interpretation and use of criteria for evaluating text quality (Eckes, 2008) or who differ in their expectation of, or involvement with, the writer (Wiseman, 2012) will rank order students' texts quite differently.

Another potential source of error in the assessment of writing is the writing task. Huang's meta-analysis (2009) showed that the two main sources of variation in performance scores are related to task characteristics. First, overall results indicated that roughly 10% of the variance is due to main task effects. This shows that average scores for performance quality differ between tasks, implying that tasks vary in level of difficulty. Second, approximately a quarter of the variance between performance scores is due to interaction effects between person and task, implying that individuals do not perform consistently across tasks.

Hence, in the context of writing performance, characteristics of raters and tasks appear to play a significant role in the assessment. Ratings of text quality are always subjective to some extent. Moreover, text quality partly depends on the topic written about and the genre written in, specified by the purpose for writing and the intended audience (Huot, 1990b). Decisions about students' writing performance are thus greatly influenced by characteristics of raters and tasks, implying that it is almost impossible to generalize to writing proficiency based on the quality of one written text, scored by one rater. Valid and reliable writing assessments should therefore include multiple tasks and raters. In addition, Lee and Kantor (2007) showed that the task facet explains more of the variability in the observed writing scores than the rater facet seems to do. Therefore, they argue, it is more efficient to increase the number of tasks in the assessment, than to increase the number of raters per task.

With this in mind, the question is how many tasks and raters are necessary for a reliable assessment of writing proficiency. Generalizability theory provides a framework for deciding upon the number of tasks and raters, given the multiple sources of measurement error (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson & Webb, 1991; for a basic introduction to this theory, see Bachman, Lynch, & Mason, 1995; Schoonen, 2005). Generalizability theory comprises a two-staged, multifaceted analysis. In the first stage, the so-called generalizability study (G-study), multiple sources of measurement error are disentangled and their variances are estimated. Based on estimates of variance components, the score generalizability can be described, reflecting the accuracy of generalizations made from the observed scores (i.e., text quality scores) to the "universe" scores (i.e., individuals' writing proficiency). In terms of generalizability theory, a universe score is the expected value of a person's observed scores over all observations to which a decision maker wants to generalize. In writing research, this universe of generalization includes all admissible raters and tasks, because one wants to generalize to writing performance across characteristics of raters and tasks (Gebril, 2009). In the second stage, the decision study (D-study), estimates of variance components obtained from the G-study are used to examine how variations in the assessment design affect score generalizability. The goal of the D-study is to choose the optimal number of tasks and raters that minimizes measurement error and consequently increases the generalizability of text quality scores to writing proficiency.

The optimal number of tasks and raters depends upon the purpose of the assessment or the decision one wants to make (Cronbach et al., 1972). In educational practices, writing scores are generally used to decide whether a student is able to write at a sufficient performance level. For absolute decisions about performance, measurement error includes both main effects of task and raters and task and/or rater effects that influence the ranking of students, that is, all interaction effects with students including random error. Writing scores are, however, at best interval scaled, rendering arbitrary means and variances of ratings (cf. Suppes & Zinnes, 1963). If means of individual tasks are arbitrary, so are differences in means between tasks. Hence, systematic variability due to tasks has no relevance, implying that writing scores reflect relative group performance rather than performance pertaining to absolute standards. For relative decisions reflecting relative group performance, only task and/or rater effects that interact with persons contribute to measurement error.

Previous research applying generalizability theory to the assessment of writing showed that at least five tasks and three raters are necessary to make valid and reliable decisions about writing skills (Coffman, 1966; Schoonen, 2005, 2012; Van den Bergh et al., 2012). Although the results of these studies seem to converge, they differ in the kind of tasks assigned to students. For instance, Van den Bergh et al. (2012) used highly similar argumentative tasks, whereas Schoonen (2005) included argumentative tasks as well as functional tasks (e.g., to give instructions, or to describe a route on a map). In some studies, tasks differed even more. Studies of Coffman (1966) and Veal & Tillman (1971), for instance, included tasks aiming at four different writing purposes, namely, describing, narrating, exposing and arguing. These tasks did also differ in the intended audience. For instance, in the study of Coffman (1966) students had to write personal essays as well as texts that were intended for another student.

Genre differences, shaped by the rhetorical situation in the writing task, are likely to lead to differences in the text that has to be produced. However, we do not know what the relationship is between genre and writing, because previous studies only included one task in multiple genres, or multiple tasks in one genre, thereby confounding the effects of genre and topic. Huot (1990b) indeed concluded in a literature review that task characteristics might elicit different writing quality, but that research on genre effects is inconclusive. In addition, Hoetker (1982) argued for a better operationalization of genre to study task variability in writing assessment.

Genre theorists and analysts, such as Swales (1990) and Bhatia (1993), have shown that texts that share the same communicative purpose and audience (i.e., texts in the same genres) are more similar in terms of their global structure, style and conventions, than texts that have a different purpose and audience (i.e., texts in different genres). This implies that the writing task does not only require content knowledge concerning the topic to write on, but also knowledge on how to conform to certain standard practices within a particular genre, in order to fulfill its communicative purpose (Bhatia, 1993). Crowhurst and Piche (1979) showed that task variables such as intended audience and mode of discourse indeed affect writing products. Their study demonstrated that there is more syntactic complexity in argument than in narration or description at both Grade 6 and 10 and even more when arguments were intended for a teacher than for a friend.

When writing products differ across genres, the process of writing may be different as well (Beauvais, Olive, & Passerault, 2011). Writing assessments consisting of highly homogeneous tasks (i.e., sharing the same communicative purpose) tend to underrepresent the construct of writing. Quellmalz, Capell and Chou (1982) claimed that writing tasks in different genres (e.g., narrative versus expository writing) tap into different cognitive processes. This claim is underpinned by Reed, Burton, and Kelly (1985) who demonstrated through the use of a secondary-task procedure that genre affects cognitive capacity during writing. Proficient writers appeared to be most cognitively engaged when writing persuasive essays and least engaged when writing descriptions. Less proficient writers, on the other hand, were most engaged when writing descriptions, but least engaged when writing narratives. Beauvais et al (2011) further showed that students adapt their writing strategy according to the genre in which they are writing. However, as was the case in earlier research, students in both studies only wrote one task in each genre. Hence, it is still unclear whether the effects are related to genre or topic. Van Weijen (2009), for instance, showed that the process of writing also differs between tasks in the same discourse mode, in this case argumentative writing.

Differences in writing products and processes suggest that performing well on one genre does not necessarily predict performance in other genres. In order to make valid inferences on writing proficiency in general, writing assessments should include multiple tasks in multiple genres. However, this will probably negatively affect the generalizability of text scores to writing proficiency, as scores based on similar tasks are more likely to correlate and thus, to generalize. Owing to the confounded effect of topic and genre in earlier research (Coffman, 1966; Reed et al., 1985; Veal & Tillman, 1971), it is still unclear whether generalization is warranted across genres, and hence, to generalize to writing proficiency in general. Therefore, the central question in the current study is whether genre has an effect on the generalizability of text quality scores.

In order to answer this question, text quality scores of different kinds of texts written by students at the end of primary education will be analyzed. These written texts constitute four different genres, differentiated by the rhetorical purpose of the writing (narrative versus argumentative writing) and audience specification. In order to disentangle genre from topic effects, students wrote multiple tasks within the same genre. Based on the magnitude of the variance components persons, raters, tasks and genre, will further be estimated how many texts, in different genres, have to be written, and how many ratings of each text are necessary in order to make inferences of differences in text quality to differences in writing proficiency.

METHOD

Participants

Written texts were obtained from 67 11-year-old and 12-year-old students in their final year of primary education (i.e., US Grade 6). These students were randomly selected from three different primary schools in the Netherlands. The participants were part of a larger research project on the effect of a digital writing program on students' writing performance (Pullens, 2012). The participants in the present study constituted the control group in this project (36% of all 186 participants), which received no explicit writing instructions and did not show a significant improvement of text quality scores during the period of study (Pullens, 2012).

Material and procedure

In total, students completed 12 writing tasks at three different moments during the study. At each moment, they received four paper-and-pencil writing tasks in four different genres. See Table 3.1 for an overview of the different tasks in four different genres, classified according to their intended purpose and audience. Narrative and argumentative writing tasks were included, as these are the two most prominent genres in both the Dutch national writing curriculum, as well as the writing standards for primary education (Meijerink, 2008).

Table 3.1 Different genres used in the present study, classified according to purpose and audience of the writing task

	Audience	
Purpose	Specified reader	Unspecified reader
Argumentative writing	Persuasive letters for a fictional company	Argumentative essays to prepare oneself for a discussion
Narrative writing	Adventure stories for readers of a school newspaper	Personal stories

Multiple tasks were collected per genre to disentangle genre effects from topic effects. Tasks within a genre were similar in terms of audience and purpose of the texts, and differed only in topic. For instance, in argumentative writing for a specified audience, students were asked to write three persuasive formal letters to fictional companies about promotional campaigns. One of these letters was to a supermarket about the collection of toys, the second letter was to a petrol station about the collection of tickets to a musical and the third letter was to a chocolate company about the collection of a music CD. See Appendix B for an overview of the specific tasks used in the present study. The format and content of persuasive letters and adventure stories in the present research were quite fixed, whereas stories about personal experiences and argumentative essays were free writing assignments.

Rating procedure

To control for the effect of handwriting on text quality ratings (McColly, 1970), the handwritten texts were retyped. Guidelines prescribed that typed texts should resemble the handwritten texts precisely, that is, all errors concerning spelling, grammar, punctuation or capitals had to be copied, as well as all modifications made by the student. Furthermore, students' names were hidden and replaced by unique codes to preserve anonymity.

For efficiency reasons, there was a design of overlapping rater teams (Van den Bergh & Eiting, 1989). Through this rating procedure, each essay was rated by a jury of three raters, without the need of having raters assess all the essays. Rater juries were selected from a total of 32 student teachers. The persuasive letters were also rated by juries of three experienced raters, who were randomly selected out of 17 teachers with at least five years of experience in the upper grades of primary education. By investigating whether ratings from inexperienced raters differ from experienced raters it was possible to analyze the effect of rater expertise on score generalizability and to determine whether results should be corrected for the raters' background.

Texts were holistically rated using benchmarks that represent the (approximate) average text quality for the writing task in question. Earlier research (Blok, 1986; Schoonen, 2005; Tillema, Van den Bergh, Rijlaarsdam, & Sanders, 2012) has shown that this rating procedure considerably increased rater reliability. After elaborate inspection of the sample of written texts, a benchmark for each writing task was selected by experienced raters. It was assured that the selected benchmarks did not contain too many grammar and spelling errors. This was important because otherwise raters' attention could be drawn away from the content and structure of the text towards mechanics. For each benchmark was explained what was average about the text according to specified criteria, such as content, structure, style, conventions and mechanics. Raters had to compare the student texts to the benchmarks, and they had to assign a score to the texts, indicating the extent to which they thought the texts were better or worse than the benchmark. As ratings are at best interval scaled (Suppes & Zinnes, 1963), each benchmark was given an arbitrary score of 100. Thus, if a rater thought a text was twice as good as the benchmark, the text was awarded a score of 200. And vice versa, a text that was considered half as good as the benchmark, according to the raters, received a score of 50. To become familiarized with the procedure, raters received a short training procedure before actually marking the essays. During this training they practiced rating several writing examples of varying quality, and discussed their judgments.

Application of a multi-group LISREL model on the covariance matrix between raters (Van den Bergh & Eiting, 1989) provides estimates of the reliability of ratings of each rater in relation to all other raters of that task (see Table 3.2). As expected, the individual reliabilities are not high (cf. Godshalk et al., 1966; Huot, 1990b; McColly, 1970). As each essay was rated by three raters, the reliability of each jury can also be estimated (see Table 3.2). These jury reliabilities are quite satisfactory, and the variance between jury's rating essays of the same task can be considered relatively low.

Table 3.2 Reliabilities of individual raters (N=32) and jury raters (N=3; 17 jurors) per writing task

Genre	Task	Reliabilities of individual raters (*ρ*, *SD*)	Reliabilities of juries (*ρ*, *SD*)
Persuasive letters	1. Collecting toys	.65 (.30)	.83 (.12)
	2. Collecting musical tickets	.64 (.25)	.83 (.09)
	3. Collecting music CD	.62 (.24)	.82 (.10)
Argumentative essays	1. Candy prohibition	.50 (.31)	.74 (.11)
	2. Smoking ban	.53 (.26)	.76 (.09)
	3. Telling tales	.58 (.24)	.79 (.09)
Adventure stories	1. Sports field	.58 (.33)	.80 (.06)
	2. Forest-fire	.57 (.24)	.78 (.11)
	3. Poison	.68 (.19)	.86 (.07)
Personal stories	1. Being frightened	.54 (.30)	.77 (.07)
	2. Being caught	.49 (.37)	.73 (.10)
	3. Home alone	.61 (.22)	.82 (.06)

Design and analysis

Since the aim of the current study is to differentiate between skilled and unskilled writers, students were the object of measurement. The other random facets included in the research design were as follows: (1) genre (g), fully crossed with persons (p); (2) tasks (t), nested within genres; and (3) raters (r), nested within tasks. This resulted in a partially nested, three-facet univariate design (p*(r:t:g)).

In total there were 2207 text scores available. Due to absence, not all students were able to complete all 12 tasks. Moreover, some texts were only rated two instead of three times. As a consequence, 205 of the total 2412 (8.5%) observations were missing, resulting in an unbalanced design.

To estimate the generalizability of writing scores, variance components were calculated for each of the seven sources of variance possible in the research design: person, genre, person by genre, task within genre, person by task within genre, raters who rated tasks within different genres, and random error. The variance components were estimated by means of the restricted maximum likelihood (REML) approach in SPSS. REML was used in order to obtain best linear unbiased estimates in unbalanced designs (Searle, 1987).

Several G-studies were performed in order to analyze the effect of genre on the generalizability of writing scores. First, to estimate the relative influence of genre, a G-study was performed in which genre was considered to be a random facet in the measurement. Second, to determine whether the stability of students' writing performance differed from genre to genre, a G-study was performed for each condition of the fixed facet genre. The expectation is that, when limiting the universe of generalization to a given genre, variance between persons will at least be as high as, or higher than, the variance for writing in general. Third, we performed an extra G-study on the ratings of persuasive letters by experienced teachers. The generalizability of these ratings was compared with the generalizability of ratings given by student teachers, to determine whether rater experience affects score generalizability.

In D-studies we approximated how many tasks and raters were needed to attain a reliable judgment about writing profiency, both within and across genres. Estimations of variance components were used to compute generalizability coefficients for relative decisions and dependability indices for absolute decisions according to varying numbers of raters and tasks. Generalizability coefficients differ from dependability indices in what is considered to be measurement error (Cronbach et al., 1972). Calculations of generalizability coefficients were based on the ratio of person variance to measurement error influencing only the ranking of persons. That is, all interaction effects with persons, specifically, interaction effects of person-by-genre, person-by-task and random error including the three-way interaction of person-by-rater-within-tasks. Dependability indices were calculated by the ratio of person variance to all sources of error including main effects of genre, tasks and raters.

RESULTS

Genre as a random facet in the measurement

The central question concerns the generalizability of text quality scores: is generalization over genres and tasks within genres warranted if students only write one text in one genre? To estimate the generalizability, the observed score variance is decomposed into seven variance components: the variance due to persons, genres, tasks within genre, raters who rated tasks within different genres, their interactions and random error. In Table 3.3, the percentages of variance associated with each of these components are summarized.

Results show that the person variance, the component of interest, only accounts for 10% of the variance in text scores. Hence, the correlation between individual text quality scores on random written texts, rated by one randomly selected rater, is low, only .32 on average. Thus, text quality scores largely (for 90%)

Table 3.3 Variance decomposition, in percentages, for persons, tasks and raters, separated by genre

Source (facet)	Percentage of variance
Person (p)	9.98
Genre (g)	11.42
Person by genre (pg)	4.01
Task within genre (t:g)	1.71
Person by task within genre (p(t:g))	19.13
Rater within task and within genre (r:t:g)	18.05
Person by rater within tasks, within genre, and error (p(r:t:g), e)	35.71

depend on facets that are not directly related to individual writing proficiency.

First of all, genre appears to be an important facet in the design. The main effect of genre accounts for 11% of the variance. Thus, average writing scores differ between genres, indicating that genres differ in difficulty; scores of text quality are slightly more similar within genres than between them. If the facet genre is not included in the analyses, the proportion of variance related to differences between persons will be overestimated. This is especially the case for decisions about writing proficiency based on absolute levels of text scores, because in these instances, genre is considered to be part of the measurement error. For instance, the observed text score of a student writing in a relative easy genre will be higher than when the same student writes a text in a more difficult genre. Decisions based on absolute writing scores are therefore affected by genre, indicating that one should write texts in more genres in order to be able to generalize to writing proficiency.

When decisions only concern the ranking of persons (i.e., relative decisions), genre is of considerably less influence: only 4% of the variance is due to the interaction of person by genre. This shows that, although the quality of a text is affected by the difficulty of a genre, better writers still outperform worse writers. This conclusion does not hold for tasks within a genre: the ranking of persons varies widely over tasks within a specific genre, as indicated by a variance component of almost 20%.

Besides the effects of genre and topic, 18% of the variance in text quality scores is explained by the interaction of rater within tasks within genre, and more than one third (35%) by random error, including the interaction of persons by rater within tasks and within genre. This residual variance is difficult to interpret, because of confounding variables in the design. It does, however, indicate that scores within and between persons fluctuate enormously, owing to differences in raters and how raters rate different tasks from different genres.

Decisions about writing proficiency across genre

In order to make an approximation about how many writing tasks and raters are needed to attain a reliable absolute judgment about writing proficiency in general, the dependability index was estimated in an absolute D-study. The dependability index is the ratio of person variance to all the variance in text scores, including unwanted variance related to all measurement facets (measurement error). The D-study showed that, in order to reach the desired level of dependability of at least .70, students should write at least four different texts in six different genres, that is, a total of 24 texts. These texts should be rated by at least three different raters. For relative decisions about writing performance, comparable levels of generalizability are attained with only 12 texts based on three different writing tasks in four different genres, rated by only two raters. Relative decisions are, however, only valid decisions when the writing assessment is used for norm-referenced testing in which the goal is to determine the relative performance of students in comparison. Relative decisions do not provide information about whether students meet a fixed standard of writing, that is, criterion-referenced testing.

All in all, the results indicate that the proportion of variance in writing scores that is explained by individual differences is rather small compared to the large effects of measurement aspects, such as rater, genre, interactions between person and genre, person and task and random error including the three-way interaction of person by rater within task.

Genre as a fixed facet in the measurement

To see whether the stability of students' writing performance differed from genre to genre, a G-study was performed for each condition of the fixed facet genre: persuasive letters, argumentative essays, adventure stories and personal stories. Moreover, for writing scores of persuasive letters, an extra G-study was performed to the ratings of experienced teachers in order to compare these ratings with ratings from student teachers. Table 3.4 summarizes the proportions of variance components (in percentages) for these G-studies.

Table 3.4 Percentage of variance due to measurement facets for different genres and rater panels differing in experience

	Teachers	Students			
	Persuasive letters	Persuasive letters	Argumentative essays	Adventure stories	Personal stories
Person (p)	17.50	14.66	12.40	12.26	24.45
Task (t)	0.27	2.33	0	3.09	1.12
Person by task (p*t)	29.60	22.63	18.35	25.92	17.39
Rater within task (r:t)	17.89	23.94	24.32	18.82	13.60
Error (p*(r:t), e)	34.73	36.45	44.94	39.90	43.44

Again, persons only accounted for a relatively small part of the variance in text quality scores (12-24%). As expected, this varies between genres; tasks in which students have to write about personal experiences show relatively more variance between students (24%) than tasks in the other genres (persuasive letters, argumentative essays or adventure stories, 12-18%).

As expected, the results do not show a main effect of tasks. The rating procedure was equal throughout all genres: every text had to be compared to a benchmark text of 100 points. For all tasks, the mean of the text quality scores across students was therefore approximately 100 points. Although scores did not vary systematically across tasks, persons performed differently on different tasks – the interaction of person by task ranged from 17% for personal stories to almost 30% for persuasive letters.

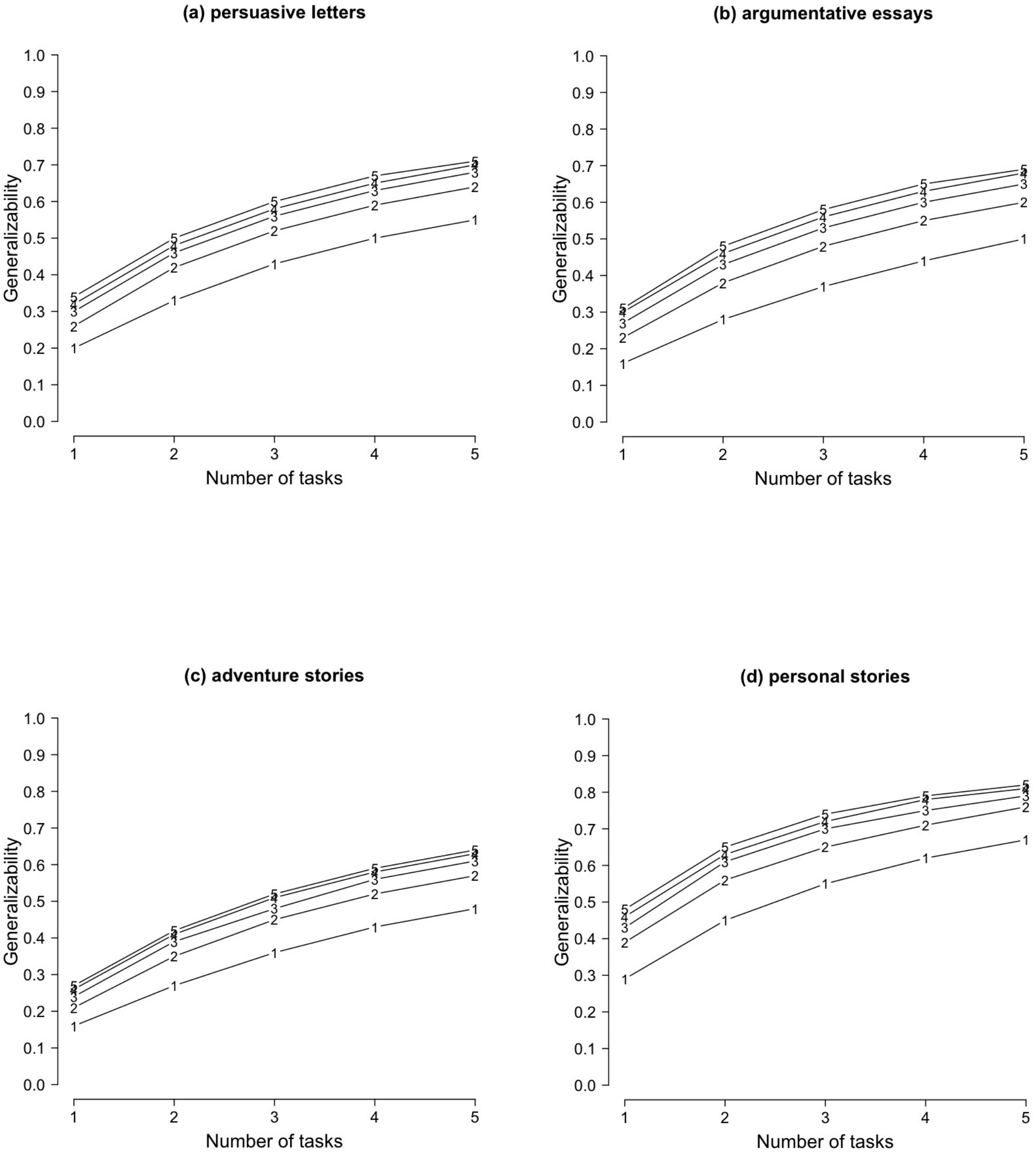
Furthermore, in line with the results presented in Table 3.4, there is a large interaction effect of rater within tasks. Judgments of raters vary from 14% for personal stories to 24% for persuasive letters and argumentative essays. This indicates that differences between raters' judgments depend on the genre of the rated texts. Specifically, raters were most familiar with the characteristics of a good personal story. To see whether consistencies in ratings were affected by rater experience, experienced teachers' ratings for persuasive letters were compared to student teachers' ratings. As expected, ratings within tasks were somewhat more consistent for experienced teachers (accounting for 18% of the variance), than for student teachers (accounting for 24% of the variance). More important, however, is the impact on generalizability, indicating that there is hardly any difference between experienced and student teachers' scores – differences in text quality are only slightly related to differences in individual writing proficiency.

Decisions about writing proficiency within specified genres

Although the stability of writing proficiency depends on the type of texts to be written, for every genre there are still other sources than students' writing proficiency at stake. The implication was that multiple tasks and multiple raters are necessary in order to generalize text quality scores to writing proficiency in a specific genre. Relative and absolute D-studies were performed to approximate the number of tasks and raters for reliable measurement of genre-specific writing.

In Figure 3.1 generalizability coefficients are plotted for multiple tasks (x-axis) and multiple raters (lines). The figures show that in order to generalize (.70 or higher) beyond given tasks or raters, one needs at least five tasks and five raters for persuasive letters, argumentative essays or adventure stories. In contrast, only three tasks and three raters are necessary for writing personal stories. Dependability indices for absolute decisions are lower in all four genres. Specifically, for five tasks and five raters, dependability indices are .66 for persuasive letters, .66 for argumentative essays, .60 for adventure stories and .80 for personal stories.

Figure 3.1 Estimated generalizability of writing scores for relative decisions, with varying number of tasks and raters within four different genres. The numbers in the lines represent the number of raters, ranging from one rater to five raters.



DISCUSSION

In the current study, aspects of the measurement of writing were disentangled in order to investigate the validity of inferences made on the basis of writing performance and to define implications for the assessment of writing. By including genre as a facet in the measurement, we obtained writing scores of 12 texts in four different genres for each student. Results indicate that across raters, tasks and genres, only 10% of the variance in writing scores is related to individual writing skill. Thus, when tasks are considered as a random selection of the universe of all possible tasks, it is quite hard to draw generalizable conclusions about writing proficiency beyond the given rater and task. More specifically, for valid and reliable inferences on relative writing performance, students should at least write three texts in each of four genres, rated by at least two raters. Even more tasks and raters are necessary for absolute decisions about writing proficiency. If it is not feasible to include different tasks in multiple genres in the writing assessment, writing scores may also be obtained by fewer but more similar kinds of tasks, that is, tasks within one genre. Absolute inferences should in this case be limited to genre-specific writing, as the results of the present study show that generalization across genre is not warranted when writing scores are obtained from texts within the same genre.

However, it should be noted that only when information is available about task compatibility or task equivalence, it is allowed to determine absolute performance levels in writing. As ratings of writing quality in writing education or writing research are generally measured on interval level, at best, differences between tasks are quite arbitrary. In this case it is not beneficial to add more tasks and raters to the assessment of writing.

Our findings emphasize the effect of genre on the generalizability of writing scores. This is in line with earlier research, showing that writing performance differs substantially between different kinds of writing tasks (Coffman, 1966; Crowhurst & Piche, 1979; Moss, Cole, & Khampalikit, 1982; Quellmalz et al., 1982; Reed et al., 1985; Schoonen, 2005; Van den Bergh et al., 2012; Veal & Tillman, 1971). However, until now, analyses of task effects were almost always based on single tasks within multiple genres (Coffman, 1966; Reed et al., 1985; Veal & Tillman, 1971) or on multiple tasks in one genre (Van den Bergh et al., 2012, Van Weijen, 2009). Hence, topic and genre effects were confounded and, as a result, inferences that could be drawn about systematic differences within and across genres were limited. The current research extends this knowledge by untangling the effects of topic and genre by including multiple tasks in multiple genres in the measurement. The results show that genre has an effect above and beyond specific task effects. This implies that if the facet genre is not included in the analyses, the proportion of variance related to differences between persons will be overestimated (see, e.g., the large person variance in Van den Bergh et al., 2012).

The results in the current study also show that the generalizability of writing scores differs from genre to genre. Presumably, for young students, it is easier to generalize to personal writing (only three texts rated by three raters are necessary) than to persuasive writing (at least five texts rated by five raters are necessary). A possible explanation for this effect is genre knowledge. When individuals are familiar with certain communicative goals, they internalize genre-specific conventions in order to reach these goals in a standardized, and thus efficient, way. As a result, writing within specific communicative events is expected to be more stable for expert members of this event (Bhatia, 1993; Swales, 1990). As young students in primary education, the object of measurement in the present study, routinely communicate about personal experiences in school, it is assumed that they have well-developed schemata for personal writing. Their genre knowledge importantly influences choices for content, structure and rhetorical style in writing. Therefore, compared to other genres, texts about personal experiences will be better comparable, and, better generalizable. Vice versa, for genres that are not practiced regularly in school, students are likely to approach each writing task as a new one, thereby limiting generalizability across tasks.

The intended audience, as specified by the writing task, may also explain the previously mentioned genre differences. Although results related to audience specification were somewhat mixed, writing performance appeared to be more stable for personal writing than for writing for specific readers. This could indicate that, at least for young students, writing about oneself is easier than writing for someone else. Bereiter and Scardamalia (1987) indeed argued that young students experience difficulties in transforming ideas and knowledge to reach a specific audience. Contrary to this theorization are the results of the quality of argumentative essays, which varied as much as the other genres. For these essays, students were asked to write down arguments in preparation for a class discussion, with the student himself as the intended audience. However, as their teacher was the leader of the subsequent class discussion, students might have written argumentative essays with the teacher as intended audience. This could have affected variability in students' writing performance, because Grade 6 students experience more difficulties when writing arguments for their teacher than writing for a friend, resulting in higher score variance (Crowhurst & Piche, 1979). Research on writing assessment is not conclusive about the relationship between components of the writing task, such as communicative purpose and intended audience, and writing quality (Huot, 1990b). It is therefore necessary to study these effects more systematically in further research.

Further, results show that writing performance differs within genre. Even when genre is included in the analysis, large variance between text quality ratings is observed, due to the interaction effect of person-by-task-within-genre. Because the estimated task effects are contaminated by effects of genre, it is not clear whether task effects are due to topic knowledge, task familiarity or due to the interaction of task-by-genre. However, it seems almost impossible to differentiate between topic and discourse mode, because choice of topic may constrain choice of public or communicative goal, and vice versa.

As expected, raters were not consistent in their judgment of text quality. This study confirms findings from previous research (Godshalk et al., 1966; Huot, 1990a) that text quality is reflected best by writing scores based on judgments of multiple raters. In line with earlier research (Gebril, 2009; Lee & Kantor, 2007) however, the effects of rater, including rater-by-task interaction, appear to be smaller than the effects of task, including genre and topic effects. This was true in all observed genres. Overall, it was estimated that the effects of task, genre and topic effects included, accounted for twice as much of the variance in writing scores as those of raters. Although the estimated residual error variance includes both rater and task effects, it seems that, ceteris paribus, it is more efficient to increase the number of tasks in writing assessment, than the number of raters.

Previous research indicated that rater experience might be an explanation for rater variability (Barkaoui, 2007; Schoonen, 2012). However, in this research, there was no effect of rater experience, at least not for the judgments of persuasive letters. Ratings of experienced teachers are more precise and consistent over tasks, given the smaller interaction effect of rater by task, but they are not superior to judgments of student teachers, as ratings of experienced teachers show larger interaction effects of person by task.

Hence, the present study shows that writing performance largely depends on the writing task. This raises the important question of how to interpret this finding? In general, and in line with generalizability theory, writing proficiency is considered to be a relatively constant disposition of a person. Writers are assumed to perform in a more or less consistent way across tasks. In terms of generalizability coefficients, this means that writing proficiency comprises only shared variance between tasks; all other score variance is considered to be measurement error. It can, however, be questioned whether task effects or the interaction of person-by-task should be regarded as measurement error. For instance, there are researchers who consider specific task effects as part of writing proficiency (Chalhoub-Deville, 2003; Read & Chapelle, 2001; Verheyden, 2011). This interactionalist view on language performance assumes that writing proficiency interacts with its context, implicating that changes in the writing task (i.e., context) will lead to changes in text quality (writing performance). This, however, has significant implications for the generalizability of the inferences made, and thus of the definition of the construct of writing proficiency (for similar reasoning, see Schoonen, 2012).

Even when writing is specified to narrower domains, writing performance is likely to fluctuate across tasks. Large task effects could, at least partly, be reduced by proper education in writing. Earlier research has shown that students do not regularly practice writing in class and that, when they do write, they hardly receive any feedback on their writing process or product (Henkens, 2010). Students therefore lack an effective approach to writing that would lead to a more consistent writing performance of higher quality. Hence, students' writing could be improved by learning effective strategies for transferring general writing principles to novel writing tasks.

Parkes (2001) has already hypothesized that task variance in performance assessment may be reduced by facilitating the transfer of knowledge across performance situations. According to his review of literature on transfer issues, effective transfer largely depends on a subtle balance between general knowledge and specific situations. He proposes three broad solutions for the transfer problem. First, students should get a good understanding of the general approaches to cognitive problem solving. In the context of writing, this means that students should know the characteristics of a good text, and how content, structure and style relate to effective writing. Second, students should form general schemata of applying general principles to specific tasks. In order to learn how to use these schemata, concrete examples of tasks should be provided. Moreover, writing strategies will help students to produce texts of a more or less consistent quality. Third, tasks should be well defined, for instance, by making the goals of the task explicit. Clear tasks promote students to see the analogy between tasks and to match the general model of earlier experiences to new situations. A writing task, therefore, should contain explicit information about the communicative goal, the topic and the intended audience. Future research should empirically test whether these solutions indeed affect the ability to transfer general writing knowledge to specific writing tasks, and thereby, reduce task variance.

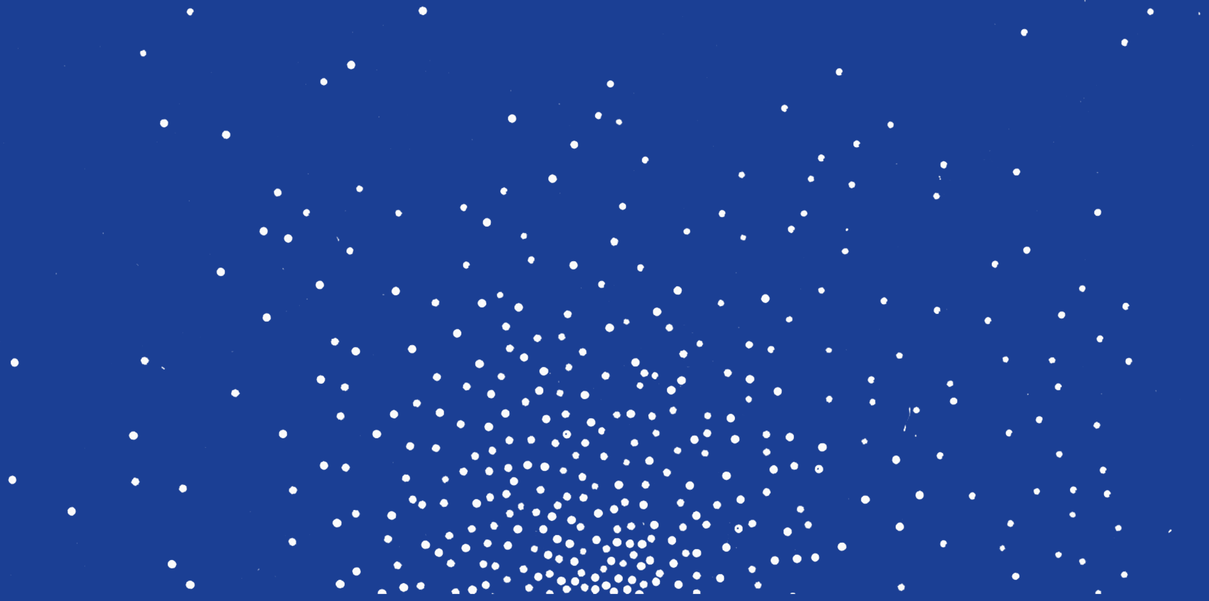
It is very likely that some other features in this study have affected the outcomes. For instance, ratings of text quality were task dependent as benchmarks differed between tasks. This does not necessarily affect the comparability of text quality scores between tasks, as text quality was measured at best on interval level. As a consequence, task variance may be underestimated, making it difficult to interpret absolute decisions. Moreover, it is possible that raters used different criteria for rating text quality per task, which could result in artificially high estimates of the interaction of person by tasks. After all, by applying different criteria to the tasks, the rank order of students might differ between tasks; good writing in one task does not necessarily imply good writing in another task. Nevertheless, jury reliabilities appeared to be acceptable for all writing tasks, suggesting that raters marked the essays in more or less the same way. It is thus more likely that rater variance contributes to random noise related to interactions between rater and text. Whereas rating criteria may vary for texts in different genres (e.g., persuasive writing is rather different from storytelling), criteria for genre-specific writing should be more or less alike. Further research should therefore use rating procedures that support raters in a more task-independent way, at least for rating texts that are similar in terms of their communicative purpose and audience. Recent research has already suggested that benchmarks can be used for different tasks within the same genre (Tillema, 2012). This is in line with the finding that benchmarks promote raters to judge texts as a whole (Schoonen, 2005).

In writing research as well as in writing education, writing proficiency is still quite often assessed with a single writing task rated by one rater. The current study shows, however, that decisions regarding writing proficiency based on one written text are not very reliable. Neither are decisions on multiple, but highly similar, texts. Because the ability to write differs from genre to genre, generalizable inferences are not appropriate. In order to draw conclusions about writing in general, writing assessment should rather include multiple tasks in multiple genres rated by multiple raters. If it is not possible to include multiple texts in different genres, for instance, because of time or money constraints, decisions should be limited to genre-specific writing.

Chapter 4

BENCHMARK RATING PROCEDURE, BEST OF BOTH WORLDS? COMPARING PROCEDURES TO RATE TEXT QUALITY IN A RELIABLE AND VALID MANNER

This chapter is based on: Bouwer, R., Koster, M., & Van den Bergh, H. (2016). Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. Manuscript is submitted for publication.



Assessing students' writing performance is essential for effective writing instruction; however, raters vary considerably in how they evaluate text quality, causing unreliability in the writing assessment. Analytical rating procedures that restrict the freedom of the rater may lead to more reliable text quality ratings, but this is often at the cost of its validity. The present research investigates a benchmark rating procedure in which raters are instructed to rate text quality holistically, while supported by benchmarks that represent the range of text quality. In two separate studies it is examined whether (a) a benchmark rating procedure leads to reliable, valid and generalizable scores, by comparing it to holistic and analytic ratings, and (b) the same benchmark scale can be used for rating texts in different topics and genres. Results indicate that benchmark ratings were associated with less rater variance than holistic ratings, and less task-specific variance than analytic ratings. Moreover, raters were able to use the same benchmark scale for rating different writing tasks with the same reliability, at least when texts are written in the same genre. Taken together, a benchmark rating procedure is a promising approach to assess students' writing performance in a reliable, valid and practical manner.

INTRODUCTION

Assessment of writing is essential for teaching and learning to write (Huot, 2002). Teachers can use writing assessments to get an impression of their students' writing performance, especially of the strengths and weaknesses of their students' writing. Further, writing assessments can be used to monitor students' progress in writing over time. Based upon the assessment of students' performance on a writing task, teachers can take informed decisions on how to improve students' writing further, for instance by adjusting their whole-class writing instruction or providing individual feedback. The more insight teachers have in students' writing performance, the better they can adapt their feedback to students' individual needs, which is essential for effective writing instruction (Bouwer, Koster, & Van den Bergh, 2016b).

Information about students' writing proficiency has to be inferred from their performance on writing tasks (Bachman & Palmer, 1996; Weigle, 2002). It is generally assumed that proficient writers write texts of higher quality than their less proficient peers. However, there are two problems hampering the assessment of text quality. First, text quality is determined by a combination of multiple features (Huot, 1990b; McColly, 1970). For instance, a text can be well written because ideas are well developed or structured in a logical way, but at the same time it can be poor because of errors in grammar, spelling, punctuation or conventions (or vice versa). Raters have to decide to what degree a text meets the required standard, taken into account the different features that underlie text quality. Second, instead of having one fixed standard of what a good text should look like, raters have their own standards of what constitutes a good text, which influences how they rate text quality and how much weight they apply to certain aspects of writing (cf. Barkaoui & Knouzi, 2012; Charney, 1984; Eckes, 2008; Meuffels, 1994). As a result, ratings vary considerably between raters, causing unreliability in the assessment of writing (Blok, 1985; Eckes, 2012; Mullis, 1984).

Reliability of writing scores

Rater variability was first demonstrated in 1961, in a pioneering study of Diederich, French and Carlton. They asked 53 raters from different professions to score the quality of 300 texts written by college freshmen on a 9-point holistic scale, without providing any external standards or criteria marking or illustrating the scale points. The average correlation between raters' scores was low, ranging from .22 for business executives to .41 for English teachers. Overall, 94% of the papers received at least seven different grades and no paper received less than five different grades. The researchers distinguished five rater types on the basis of raters' primary focus during the rating process, which were either on ideas, form, style, mechanics or the wording of a text.

Rater inconsistency occurs not only between raters, but also within raters. Individual raters tend to assign different scores to the same text on different occasions (cf. Breland, 1983; Coffman, 1971). It is evident that both types of rater variability are problematic for the assessment of writing performance. Based on the ratings of one rater it is hardly possible to generalize to students' writing proficiency, as the ratings may have been different if students' texts were scored by another rater or at another time. One solution to the rater problem is to have texts rated by multiple raters (Godshalk, Swineford, & Coffman, 1966). When scores of multiple independent raters are combined into one jury score they reflect writing performance more validly and reliably. Research has demonstrated that at least three to five raters are necessary for a reliable assessment of writing performance (Bouwer, Béguin, Sanders, & Van den Bergh, 2015; Schoonen, 2005). In practice, however, it is not always feasible to include multiple raters in the rating process. Moreover, when too many raters are involved who provide judgments of text quality with their own standards in mind, the ultimate text quality ratings may become trivial (Meuffels, 1994). It is therefore important to seek for alternative ways to reduce rater variability.

Rating procedures

One way to reduce rater variability is to implement a rating procedure that offers instructions or support to raters on how to rate text quality. Rating procedures vary in the extent to which the freedom of the rater is limited (cf. Wesdorp, 1981) and can roughly be summarized into two broad streams: holistic or analytic. In the following paragraphs we will explain and compare holistic and analytic rating procedures.

Holistic ratings. Holistic ratings represent raters' impression of the quality of a text as a whole. In this rating procedure, raters have to assign text quality scores according to predefined rating criteria, but it is not explicitly prescribed how the quality for each criterion has to be determined or how evaluations of specific text features should be combined into one overall evaluation of the quality of the text as a whole (cf. Charney, 1984). Thus, a holistic rating procedure limits the freedom of raters only with regards to the kind of features that they are assumed to be taking into account when rating text quality. As a consequence, rater variability will be reduced only to a limited extend (cf. Wesdorp, 1981).

Research has demonstrated that even when raters pay attention to the same criteria, their scores may vary considerably. For instance, raters may unintentionally weight certain aspects of writing over others (cf. Olinghouse, Santangelo, & Wilson, 2012). Raters may also differ in their personal preferences for certain characteristics of the writer or the written text, such as students' handwriting, which might affect their writing scores on a product-by-product basis (Barkaoui & Knouzi, 2012). This is commonly referred to as halo effects (cf. Meuffels, 1994; Myford & Wolfe, 2003; Rijlaarsdam et al., 2011; Wiseman, 2012). Furthermore, raters may differ in how they distribute scores throughout the score scale (cf. Coffman, 1971; Meuffels, 1994; Myford & Wolfe, 2003;

Rijlaarsdam et al., 2011). For instance, when raters are more severe and apply scores consistently lower than other raters, or when raters have a central tendency to rate near the scale midpoint, avoiding extremely high or low scores, raters are not in absolute agreement about students' writing performance (cf. Myford & Wolfe, 2003). However, in most of the cases such rater effects only cause systematic differences between raters, which can be considered to be irrelevant as text quality is measured at best on interval level, rendering arbitrary means and variances of ratings.

Analytic ratings. Instead of forming a holistic impression of text quality, raters can also evaluate text quality analytically by rating specific text features (e.g., separate ratings for content, structure or mechanics). An analytic rating procedure is less free than a holistic procedure as it not only limits the text features that raters have to consider while evaluating text quality, but also controls the amount of weight raters give to each feature (cf. Barkaoui, 2011). Although specific text features might be rated in a holistic manner (i.e., by taking the text as a whole into consideration), when separately scored features are combined into one composite rating for text quality, the procedure is considered to be analytic (cf. Jonsson & Svingby, 2007; Weigle, 2002).

There are various analytic scales that can be used for scoring text quality analytically, which vary in how task-specific and detailed the scoring instructions are. For instance, an analytic scale can include a list of predefined criteria that have to be evaluated one by one. It can further distinguish between different performance levels for each criterion, which is generally the case in analytic rubrics (cf. Huot, 1990b; Jonsson & Svingby, 2007). An analytic scale can also take the form of a dichotomous scale, in which a series of statements are explicated about specific text features (Cooper, 1977). This simplifies the rating task even more as the rater only has to decide for each statement whether a text includes the particular feature or not. Whereas a list of criteria can be applied to different writing tasks, a checklist with statements about the presence of certain elements in a text are highly task-specific by nature.

It is generally assumed that restriction of the freedom of the rater in analytic rating will increase inter-rater reliability. Follman and Anderson (1967) proposed that analytic rating procedures especially help raters with heterogeneous backgrounds to evaluate text quality in the same way. Breland (1983) compared different studies with either holistic or analytic scoring methods and showed that analytic scoring generally resulted in improved reliabilities. Barkaoui (2007; 2011), however, demonstrated the opposite: holistic ratings were associated with higher reliability coefficients than analytic ratings, particularly for experienced raters. Although an analytic rating procedure supported novice raters to be more self-consistent, it did not increase agreement between raters. Based on these findings he concluded that detailed rating scales are no guarantee for high inter-rater agreement, unless rating criteria are clarified and discussed in training sessions. Lumley (2002) also showed that analytic scales do not cover all problems in the rating process. Based on thinking-aloud protocols he showed that especially higher order criteria such as the relevance or clarity of ideas were hard to score, even when they were fully specified in a rubric. Raters applied scale descriptions not always in the same manner. Moreover, sometimes they even used their overall impression of the text to rate separate features, which led to high correlations among different text features. This latter finding was also demonstrated by De Glopper (1988), who showed that analytic ratings of texts written by 9th grade students largely converged to one general factor. It can therefore be questioned whether text features actually are independent constructs, and hence, whether raters are able to rate text quality in an analytic manner (De Glopper, 1988; McColly, 1970).

Thus, although analytic rating procedures are supposed to control for the variation between raters, it appears that this advantage over holistic rating procedure is rather limited. However, the process of analytic ratings is much more time consuming than evaluating text quality holistically (Huot, 1990b; Weigle, 2002; Wesdorp, 1981). Whereas for holistic ratings a text has to be read only once to form a general impression, for analytic ratings the rater has to read a text multiple times in order to come to multiple scores for different features.

Validity of writing scores

Although reliability is a prerequisite for text quality scores to be valid, it is not a guarantee (cf. Huot, 2002). For instance, counting words can be done very precisely and reliably, however, the results hardly relate to the quality of the text, as it does not convey essential information about attainment of the rhetorical goal of a text (cf. Lloyd-Jones, 1977). This may also hold for analytic ratings, as the overall quality of a text might be different from the sum of its parts (Cooper, 1977). Moss, Cole and Khampalikit (1982) demonstrated that correlations between holistic and analytic ratings, when corrected for the unreliability of the measurement, were rather low, ranging from .48 to .72. These correlations correspond to a shared variance between raters ranging from .23 to .52. This limited overlap between holistic and analytic measures of text quality is confirmed in other studies as well (e.g., Grabowski, Becker-Mrotzek, Knopp, Jost, & Weinzierl, 2014; Olinghouse et al., 2012). Hence, holistic and analytic ratings at best reflect different aspects of text quality.

Thus, it is important to know which of the two rating procedures, holistic or analytic, leads to more valid evaluations of text quality. Studies on the generalizability of text quality scores have shown that it is harder to make generalizations about students' writing proficiency based on analytic scores than on holistic scores (Barkaoui, 2007; Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). Although analytic scores may lead to more precise and reliable judgments (i.e., less rater variance), this seems to be at the cost of larger task-specific variance. This indicates a tension between reliability and validity demands in the measurement of writing (Huot, 2002; Wesdorp, 1981).

Benchmark rating procedure

To summarize, although holistic ratings are a more valid representation of the quality of a text as a whole, raters need support in using the full range of scores in a more or less equal manner in order to increase the reliability of the ratings. A way to reduce rater variance, instead of breaking the rating process down into rating separate text features, is to implement a benchmark rating procedure. In this procedure, raters receive benchmarks that illustrate the points on a rating scale, representing the range of text quality that can occur in a given group (Coffman, 1971; Cooper, 1977; Mullis, 1984; Schoonen, 2005; Wesdorp, 1981). For each benchmark it is described why it does or does not meet the rhetorical goal (Lloyd-Jones, 1977). Raters have to compare students' texts with the benchmarks on the scale to rate a text accordingly. As a benchmark rating procedure facilitates raters to rate the quality of a text holistically, but also offers support by providing a standard of text quality in which different text features are operationalized and performance levels are distinguished, the ratings may be valid as well as reliable, and hence, be considered the best of both worlds (Schoonen, 2005; Wesdorp, 1981).

Benchmark rating procedures have been widely used in previous studies to evaluate students' writing performance (Bouwer, Koster, & Van den Bergh, 2016a; De Glopper, 1988; De Smedt, Van Keer, & Merchie, 2015; Rietdijk, Van Weijen, Janssen, Van den Bergh, & Rijlaarsdam, 2016; Tillema, Van den Bergh, Rijlaarsdam, & Sanders, 2012). Yet, there is hardly any research in which benchmark ratings are compared to holistic or analytic ratings regarding their reliability and validity. Research from Schoonen (2005) indicated smaller task-specific variance for benchmark ratings compared to analytic ratings. Thus, benchmark ratings seem to be better generalizable across tasks than analytic ratings. Further, research revealed that raters evaluate text quality more reliable when they have to compare two texts with each other, than when they have to assign a score to single text (Pollitt, 2012; Thurstone, 1927). Texts can be compared to each other (Lesterhuis, Verhavert, Coertjens, Donche, & De Maeyer, 2015) or against some fixed example texts (i.e., benchmarks) that represent the range in text quality (Blok, 1986). The advantage of the latter option is that raters will be less likely to adapt their standards during the rating process as students' texts are always compared to the same benchmarks. As a result, fewer comparisons have to be made in order to reach reliable scores.

However, a disadvantage of using a scale with fixed examples may be that for each writing task a separate set of benchmarks has to be selected. This limitation is emphasized by several researches, who hypothesized that a benchmark rating procedure can only support raters when the to-be-rated texts are similar to the benchmarks on the scale in terms of writing topic and genre (Feenstra, 2014; Meuffels, 1994; Pollmann, Prenger, & De Glopper, 2012). This could impede the large-scale use of benchmark rating scales, as the development of a benchmark scale takes time and expertise (Feenstra, 2014; Pollmann et al., 2012). In contrast, Tillema et al. (2012) showed that examples offer enough support to rate text quality of different writing tasks in a reliable manner, even when the topics (and language) are different. This may suggest that a benchmark scale can be used for the evaluation of texts that slightly differ from the benchmarks.

Aim of the study

To summarize, a benchmark rating procedure seems to be a promising approach to assess students' writing performance, but it has to be examined whether the reliability, validity and generalizability of benchmark ratings are different from holistic or analytic rating procedures. Further, in order to understand the usefulness of a benchmark rating procedure, it should be empirically investigated whether it is possible to use the same benchmark scale in a reliable way for rating texts from tasks that differ from the benchmarks on the rating scale in topic and genre. In two separate studies, these research questions were investigated. We hypothesized that benchmark ratings are more reliable than holistic ratings, as raters have more support during the rating process. Moreover, we expected that benchmark ratings are better generalizable to students' writing proficiency than analytic ratings and that the ratings converge more with holistic than with analytic ratings. Finally, we tested whether the similarity in topic and genre between the benchmarks and to-be-rated texts influenced the reliability of benchmark ratings.

STUDY 1

The central question in study 1 was whether a rating procedure with benchmarks leads to more reliable, valid and generalizable scores of text quality than a holistic or analytic rating procedure.

Method

Participants and procedure. A total of 36 undergraduate students (31 female, 5 male) participated in this study. The students were from the Department of Language, Literature and Communication. Mean age of participants was 22.62 years ($SD = 3.41$). Participation was voluntarily; students did not receive any reward. Participants were randomly assigned to one of two text samples (Yummy or Smurfs) and one of three rating procedures (holistic, benchmark or analytic). Participants were instructed to rate the quality of the texts using the assigned procedure and to write down their scores on the text. They were instructed only to change scores in exceptional cases. The ratings were done independently in class and took about one hour. Some participants completed the ratings faster than others, but they had to wait until the other participants had finished.

Text samples. Texts were randomly selected from an earlier research project on students' writing performance in grade 6 (Pullens, 2012). We used texts from two persuasive writing tasks in which students were asked to write a formal letter to a fictional company about a problem with a promotion campaign. One task was about collecting points on Yummy Yummy candy bars and the other task was about collecting Smurfs in the supermarket, see Appendices C and D for task descriptions. From the total of 186 participating students, 99 students completed both writing tasks. We randomly selected one-third of these students ($N = 34$). This selection included low, average as well as high proficient writers, which resulted in two samples of 34 texts reflecting the range of writing performance of students in grade 6. The handwritten texts of the students were retyped including all errors concerning spelling, grammar, punctuation or capitals, as well as all modifications made by the student.

Holistic rating procedure. Participants who were assigned to the holistic rating procedure received written instructions to rate each student text on a scale from 1 to 10 based on their general impression of the quality of the text. They were instructed to provide these holistic ratings with the following five criteria in mind: (1) content: the extent to which a text includes a clear problem statement and a question to the reader, (2) structure: the extent to which a text is coherent and understandable, (3) style: the extent to which the language is effective and varied and fits the audience and purpose, (4) conventions: the extent to which the text meets the formal requirements with regards to form, such as whether the letter includes contact details, a date and a proper salutation and closure, (5) mechanical aspects: whether the grammar, spelling, and punctuation is sufficient. Raters were free in how they weighted these features into an overall score for text quality.

Benchmark rating procedure. Participants who were assigned to the benchmark rating procedure received a continuous rating scale with five benchmarks that represent the range of writing quality of students in grade 6. They were instructed, on paper, to rate the quality of each student text holistically by comparing the text to the benchmarks on the scale in order to position it on the scale and to rate it accordingly. There were two benchmark scales, one for each writing task. The scale can be considered as an interval scale. A benchmark of average quality marks the center position on the rating scale, and is given an arbitrary score of 100 points. The other benchmarks on the scale are one (115 points) and two (130 points) standard deviations above average, and one (85 points) and two (70 points) standard deviations below average. Raters' scores could be either within or outside the range of the benchmarks.

The texts that were selected as benchmarks originated from the same research project as the text samples in the present study (Pullens, 2012). In this project, text quality scores were given by juries of three independent raters who rated the texts holistically ($\rho_{average} = .77$). The text quality ratings were normally distributed. Benchmarks were selected based on two criteria: first, the text had to be a good representative of the scale point (-2 SD, -1 SD, 0, +1 SD, +2 SD), and second, there had to be high agreement between the three raters who rated the quality of the text (i.e., low variance between raters' scores). For each benchmark was described why the text was representative for its location on the scale, using the same criteria as the holistic rating. See Appendix I for an example of the benchmark rating scale for the Yummy task.

Analytic rating procedure. Participants assigned to the analytic rating procedure received a scoring guide consisting of 15 items on the presence or absence of specific features in the text. This scoring guide was developed and used earlier by the Dutch Institute of Educational Measurement to evaluate students' level of writing proficiency (Kuhlemeijer, Van Til, Hemker, De Klijn, & Feenstra, 2013). There were 6 items about the content of the text (e.g., does the student state something about the eight points already collected?), 6 items were about the structure and conventions of the text (e.g., is there a formal salutation?) and 3 items were related to the communicative goal and the audience (e.g., are the arguments convincing?). An example of the analytic scoring guide for the Yummy task is included in Appendix H. Students had to score each item with one point if the particular feature was present in the text and with zero points if it was absent. The total score for text quality was based on the 15 items in the scoring form. Hence, scores ranged from 0 to 15.

Data analysis. For each task and rating procedure we examined the inter-rater reliability, the construct validity and the generalizability of the writing scores. The inter-rater reliability was estimated by Cronbach's alpha, indicating the consistency of ratings from independent raters. A high coefficient value indicates that raters give high and low scores in a similar pattern. The acceptable size of a reliability coefficient depends on the purpose of the measure. For research purposes, coefficient values of 0.6 - 0.7 are considered to be an acceptable level of rater agreement and values of 0.8 or higher indicate good rater agreement (Stemler, 2004). Because Cronbach's alpha increases as the number of raters does, we used the Spearman-Brown formula[1] to determine the reliability coefficients for an equal number of raters for each rating procedure. Further, we applied a K-sample significance test to compare the reliability coefficients of text scores within the same sample (Feldt, 1980; Hakstian & Whalen, 1976).

The construct validity of each rating procedure is estimated by correlating the scores of each rating procedure with the scores obtained from the other two rating procedures within the same task (Cook & Campbell, 1979). This is commonly referred to as convergence validity: measures of a similar construct (i.e., text quality) should converge. Thus, high correlation coefficients between

[1] When k is the factor by which the length of the test is changed, and r_x is the reliability of the original test, the reliability of r_{kk} can be estimated by: $r_k = \frac{k * r_x}{[1+(k-1)* r_x]}$

rating procedures indicate that they measure the same construct. As all three rating procedures in the present study are developed to measure the construct of text quality, it is assumed that they will highly correlate with each other. Low correlations indicate a threat to construct validity, indicating that at least one of the rating procedures measures something else (Cook & Campbell, 1979). Since the rating procedures will not have a perfect reliability due to measurement error, correlations between scores from two rating procedures will suffer from attenuation. Therefore, we corrected for attenuation attributable to unreliability by dividing the observed correlation coefficient by the product of the square roots of the two relevant reliability coefficients (Lord & Novick, 1968). These attenuated correlation coefficients reflect the true correlations between rating procedures.

We used the framework of generalizability theory to estimate the generalizability of writing scores for each rating procedure (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972). First, we disentangled different sources of variation in the measurement of writing. In the present study the sources of variation included the writer, task, writer-by-task interaction, rater-by-task interaction, and the three-way interaction between writer, task and rater, including random error. Because raters were nested within writing tasks, the estimation of rater effects is contaminated by the rater-by-task interaction. Estimated variance components were computed for each source of variation using SPSS, separately for each rating procedure. Second, we approximated the generalizability coefficient for each rating procedure, by estimating the variance that is associated with the writer (i.e., true score) as a proportion of the total amount of variance (including all sources that are regarded as measurement error). Hence, generalizability coefficients indicate the extent to which text quality scores can be generalized to students' writing proficiency. Third, we compared the generalizability coefficients, in order to determine whether the generalizability of writing scores depends on the rating procedure.

Results

Reliability estimates. The means and standard deviations for holistic, benchmark, and analytic rating procedures per writing task are presented in Table 4.1. It is shown that the average scores for the Smurf task are somewhat lower than the average scores for the Yummy task. Table 4.1 also presents the interrater reliability coefficients per writing task and rating procedure. Results show that the Cronbach's alpha reliability coefficient of mean scores ranged from .79 to .93, depending on the rating procedure and writing task. A reliability coefficient of .93 indicates that ratings will overlap for 86% with scores given by another sample with the same number of raters.

However, as the number of raters was not the same for each rating procedure, we used the Spearman-Brown formula to estimate reliability coefficients for scores based on one, two or three raters. These reliability coefficients are presented in Table 4.2. For the Yummy task, the reliability of benchmark ratings for a single rater was .64, which was not significantly higher than the reliability of holistic rating, which was .48 ($F(1, 33) = 1.44$, $p = .15$). The reliability of benchmark ratings equaled the reliability of analytic ratings ($F(1, 33) = 1.00$, $\mathrm{p} = .50$). For the Smurf task, the results were the same: there were no significant differences between the reliabilities between benchmark and holistic ratings (.48 versus .60, $F(1, 33) = 1.30$, $p = .23$) or between benchmark and analytic ratings (.48 and .54, $F(1, 33) = 1.13$, $p = .36$).

Further, the results demonstrated that the inter-rater reliability is higher when ratings are based upon judgments of multiple raters than when ratings are based upon a single rater. For none of the rating procedures a desired reliability level of .70 is reached when there is only one rater involved in the measurement. At least two or more raters are necessary to reach a sufficient level of reliability.

Table 4.1 Means, standard deviations and number of raters for holistic, benchmark and analytic rating procedure

		Yummy				Smurf			
Rating procedure	Scale	*N*	Mean	*SD*	α	*N*	Mean	*SD*	α
Holistic	0 - 10	4	5.93	1.67	.79	7	4.94	1.86	.91
Benchmark	0 - ∞	7	91.00	20.48	.93	5	82.59	19.88	.82
Analytic	0 - 15	7	8.76	3.01	.93	6	8.30	2.72	.87

Table 4.2 Average reliability coefficients per rating procedure and writing task

	Yummy			Smurf		
	Reliability (*ρ*)			Reliability (*ρ*)		
Rating procedure	1 rater	2 raters	3 raters	1 rater	2 raters	3 raters
Holistic scale	.48	.65	.74	.60	.75	.82
Benchmark scale	.64	.78	.84	.48	.65	.74
Analytic scale	.64	.78	.84	.54	.70	.78

Convergent validity. Table 4.3 shows the (attenuated) correlation coefficients between the three rating procedures, separated by writing task. It is shown that the intercorrelations for all combinations are high, but that there are some differences between the writing tasks. For Yummy, the average attenuated correlation between holistic, benchmark and analytic ratings is .98 and for Smurf .79. Further, it is shown that the correlations between benchmark and holistic scores were higher than the correlations between benchmark and analytic scores. This was even stronger for the Smurf task, compared to the Yummy task. Together, these results demonstrate that benchmark ratings are converging to the same construct as holistic ratings, whereas there was less convergence between benchmark and analytic ratings, especially for the Smurf task.

Table 4.3 Correlation coefficients between holistic, benchmark, and analytic ratings for the two writing tasks

	Yummy			Smurf		
	Holistic	Benchmarks	Analytic	Holistic	Benchmarks	Analytic
Holistic	–	.90	.83	–	.83	.63
Benchmark	1.00	–	.86	.96	–	.62
Analytic	.97	.93	–	.86	.73	–

Note. Uncorrected correlations are above the diagonal and attenuated correlations are below the diagonal.

Table 4.4 Variance components as proportions of the total variance estimated per rating procedure

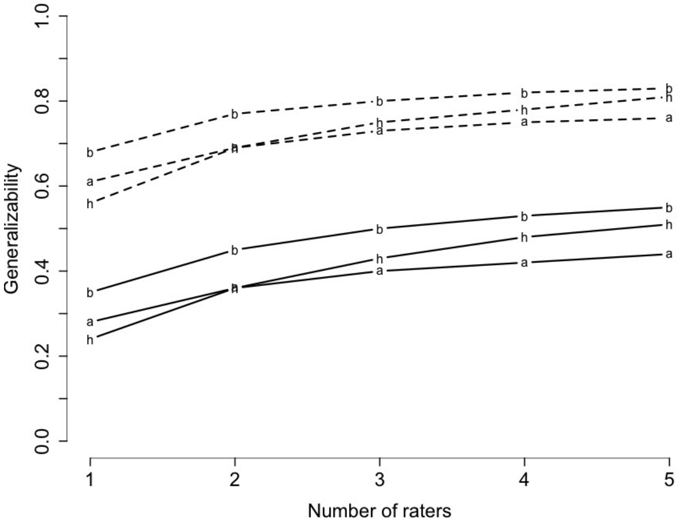
Source	Holistic	Benchmarks	Analytic
Student (s)	.23	.33	.28
Task (t)	.06	.06	.00
Student by task (st)	.10	.18	.27
Rater within task (r:t)	.35	.08	.08
Student by rater within tasks, and error (s(r:t),e)	.26	.35	.37

Generalizability estimates. Table 4.4 shows for each rating procedure the proportion of variance that is associated with the writer, task, writer by task interaction, rater by task interaction, and the interaction between writer, tasks, and raters, including random error. Results show that the percentage of variance that is related to the writer was higher for the benchmark rating condition (33%) than for the holistic or analytic rating condition (respectively 23% and 28%). So, for benchmark ratings, 33% of the variance in text quality scores can be explained by the ability of the writer. Hence, the expected correlation between two random written texts, rated by one randomly selected rater who is supported by benchmarks will equal $\surd.33$, which is .57. The expected correlation between two texts written by the same writer will be lower when the rater evaluates the text holistically without benchmarks ($\surd.23 = .48$) or analytically ($\surd.28 = .53$).

The variance components in Table 4.4 further show that holistic ratings included large rater within task variance (.35). This indicates that a holistic rating procedure yields relatively rater-specific judgments, which makes generalization across raters rather difficult. Differences between raters were minimized for benchmark and analytic ratings (both .08). It was also demonstrated that analytic ratings included large student by task interaction (.27), compared to holistic and benchmark ratings (respectively .10 and .18). This indicates that an analytic rating procedure yields relatively task-specific judgments, which makes generalization across tasks more difficult.

Based upon the estimated variance components it is possible to make an approximation of how many writing tasks and raters are needed to make generalizations about students' writing performance. This approximation reflects the generalizability coefficient, which is estimated as the proportion of student variance to the sum of all variance components including student variance. The generalizability coefficient takes the relative weight of the variance components into account, which depends on the number of tasks and raters included in the writing assessment. Figure 4.1 shows the generalizability coefficients for the three rating procedures and varying number of tasks and raters. In the figure it is shown that although benchmark ratings are more generalizable than analytic or holistic ratings, a desired level of generalizability of at least .70 is not reached when the assessment includes only one writing task, regardless the number of raters. The generalizability increases when more tasks are included. For benchmark ratings, a desired level of generalizability of at least .70 is reached when students write at least four tasks, measured by at least two different raters. More raters (at least three) are needed when text quality is rated in a holistic or analytic manner.

Figure 4.1 Estimated generalizability of writing scores for one task (solid lines) and four tasks (dashed lines), with varying number of raters. Lines represent the rating procedure that is used by raters, b: benchmark rating procedure, a: analytic rating procedure, h: holistic rating procedure.



Conclusion

Based on the results we could not establish a difference between the three rating procedures, indicating that benchmark ratings are as reliable as analytic and holistic ratings. From both the reliability and generalizability analyses we can conclude that multiple raters are necessary to establish a sufficient level of reliability, regardless of the rating procedure that is used. It is also necessary to include multiple tasks in the writing assessment in order to account for the large task effects that make generalizations to students' individual writing proficiency difficult. However, in general, less tasks and raters are needed when raters use a benchmark rating procedure to rate text quality, instead of a holistic or analytic rating procedure. Further, validity measures have shown that benchmark ratings are highly convergent with holistic ratings, indicating that they measure the same underlying construct. Correlations between benchmark and analytics ratings were lower than correlations between benchmark and holistic ratings. Taken together, the results show that a benchmark procedure leads to both reliable and valid ratings.

STUDY 2

In this study we examined whether the same benchmark scale can be used for rating different kinds of texts. Therefore, we compared the inter-rater reliability of text quality ratings for texts that (a) were similar to the benchmarks on the scale regarding both topic and genre, (b) differed in topic, but were in the same genre as the benchmarks, or (c) differed in both topic and genre.

Method

Participants. Ten undergraduate students (9 female, 1 male) participated as raters in this study. The students were from the Department of Language, Literature and Communication. Mean age of participants was 21.90 years ($SD = 1.73$). None of the participants indicated to have experience with rating the quality of texts written by students in elementary grades. Participation was voluntarily.

Materials and procedure. Raters rated the quality of a sample of 120 texts written by students in grade 6 from different elementary schools. The sample consisted of texts from three different writing tasks, written in earlier research projects on elementary students' writing development (Koster, Bouwer, & Van den Bergh, 2016a; Pullens, 2012). Per task we randomly selected 40 texts, resulting in a total sample of 120 texts reflecting the range of writing performance of students in grade 6. For two writing tasks, students had to write a persuasive letter to a fictional company about a problem with a promotion campaign (Pullens, 2012). These writing tasks were similar with regard to genre and differed only in topic. One task (Yummy) was about collecting Yummy Yummy candy bars for a music CD. The other task (Smurf) was about collecting Smurfs for a digital camera. Task descriptions are included in Appendices C and D. The third writing task (Like) differed from the other tasks in topic as well as in genre: students had to write a letter of advice to a fictitious peer on how to get good grades for writing (Koster et al., 2016a). A description of this task is included in Appendix E.

A benchmark rating procedure was used to rate the quality of the 120 student texts. In this rating procedure, text quality is rated holistically by comparing each student text to a continuous scale with five benchmarks that mark different quality levels (-2 SD, -1 SD, 0, +1 SD, +2 SD). Based upon this comparison, raters can locate the text on the scale and score it accordingly. Scores could range either within or outside the range of the benchmarks. Two different benchmark scales were used in the present study, one with benchmarks from the Yummy task and one with benchmarks from the Smurf task. The benchmark rating scales were the same as in Study 1; see Appendix I for an example.

Half of the raters used the Yummy benchmark scale to rate the quality of the texts and the other half of the raters used the Smurf benchmark scale. For both groups the order of 120 student texts was as follows: raters first received the sample of texts that was similar to the benchmarks in both topic and genre; second, the texts that were similar to the benchmarks in genre but differed in topic; and third, the texts that were different to the benchmarks in both genre and topic. For example, the raters with the Yummy benchmark scale first had to rate they Yummy texts, then the Smurf texts and finally the Like texts. The order of texts within samples was kept the same for all raters.

All raters received written instructions and a short training of 15 minutes before starting with rating the quality of the texts. The Yummy and Smurf groups were trained separately. Training involved discussion of the benchmarks and guidelines of the rating procedure. Raters were instructed to work independently, and to focus primarily on the communicative effectiveness of the text while comparing it to the benchmarks on the rating scale. They practiced scoring of three selected example texts, one of poor quality, one of average quality and one of high quality. The example texts were from the same task as the benchmarks on the scale. Further, it was explicated that the texts had to be rated one by one, and that they were not allowed to change or adjust their ratings once they were given. They had a maximum of 120 minutes to rate all texts. Raters were able to get a coffee in between, or to go to the restroom, but were instructed to do this without discussing the ratings or the rating procedure. After the ratings, participants were asked to indicate on a 5-point Likert scale how they perceived the difficulty of the rating procedure for the three different samples of texts.

Data analysis. There were two groups of five raters who each rated 120 texts. Inter-rater reliabilities for the three text samples were estimated per rater group, using Cronbach's alpha. As the number of raters was equal in both groups, the alpha coefficients can be compared with each other. A K-sample significance test was used to compare the Cronbach's alpha reliability coefficients within groups (Feldt, 1980; Hakstian & Whalen, 1976).

Results

The means, standard deviations and inter-rater reliability for the three writing samples and two groups are presented in Table 4.5. For the Yummy rater group, results show no significant differences in the reliability coefficients for ratings of texts that were similar to the benchmarks in both topic and genre versus texts that were only similar to benchmarks with regard to genre (.92 and .90, $F(1, 39) = 1.21$, $p = .27$), or versus texts that were different from the benchmarks in both genre and topic (.92 and .87, $F(1, 39) = 1.52$, $p = .10$). Hence, raters' scores were highly consistent for texts that were very similar to the benchmarks as well as for texts that were very different from the benchmarks, indicating that a benchmark rating scale can be used for different tasks.

Table 4.5 Number of raters, means, standard deviations and inter-rater reliability for two rater groups and three writing samples

	Yummy rater group				Smurf rater group			
Similarity	*N*	Mean	*SD*	α	*N*	Mean	*SD*	α
Topic + genre	5	89.87	18.18	.92	5	94.09	14.53	.91
Genre	5	90.50	17.44	.90	5	88.85	17.81	.92
None	5	90.83	15.95	.87	5	88.43	15.97	.84

Results for the Smurf rater group were slightly different. Whereas the reliability coefficient of ratings for text in the similar genre were not significantly different (.91 and .92, $F(1, 39) = 1.23$, $p = .26$), the reliability for ratings from a different type of text were significantly lower than the reliability of the texts that were in the same genre as the benchmarks (.84 versus .92, $F(1, 79) = 1.81$, $p < .01$). Hence, raters' scores were highly consistent for texts that were in the same genre but less consistent for texts that were from a different genre than the benchmarks on the scale.

Questionnaire data revealed that raters experienced more difficulties when they had to rate texts in another genre and topic ($p < .01$) than when they had to rate texts that were more similar to the benchmarks on the rating scale (respectively $p = .40$ and .55), $F(2, 27) = 46.02$, $p < .01$.

Conclusion

Taken together, the results suggest that a benchmark rating procedure can be used for rating texts from different tasks, at least when the tasks are within the same genre. It seems that benchmarks of a different genre provide less support to raters to rate text quality in a reliable way.

GENERAL DISCUSSION

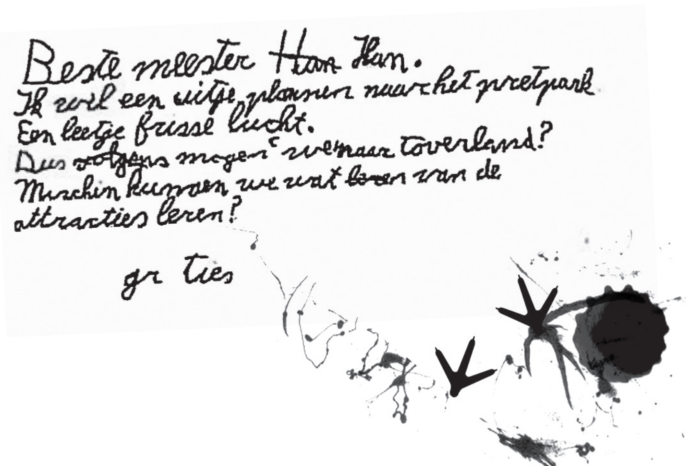
The present study examined, in two separate experiments, the reliability, validity, and usability of a benchmark rating procedure. The results of the first experiment favor a benchmark rating procedure over a holistic or analytic rating procedure. Compared to analytic ratings, benchmark ratings were not more reliable, but they were more generalizable because benchmark ratings included less task-specific variance. Further, correlations between analytic and benchmark ratings were lower than correlations between holistic and benchmark ratings, indicating that the latter two rating procedures are converging to the same construct. With regards to the comparison between benchmark and holistic ratings, the results were contrary to expectations. Although we expected that raters would have experienced more support from a benchmark rating scale than from a list of criteria that had to be evaluated holistically, there were no significant differences in the inter-rater reliability between the two rating procedures. The variance component analysis showed, however, that benchmark ratings included less rater variance than holistic ratings. This implies that fewer raters are needed for benchmark ratings to reach a sufficient level of generalizability, compared to holistic ratings. Based on these results it can be concluded that the benchmark rating procedure is a promising approach to assess students' writing performance in both a reliable and valid manner.

Further, the results have shown that benchmark scales can be used for rating different writing tasks, at least when texts are written in the same genre. Although the order in which texts were rated was confounded with the similarity between the texts and benchmarks (i.e., raters always rated more similar texts before they rated different texts, which might have affected the outcomes), we were able to replicate the findings for two different benchmark scales. Thus, based on the present results it does not seem necessary that benchmarks are from the same writing task with the same topic as the texts that have to be rated. This result does not confirm previously raised concerns on the use of benchmark scales (Feenstra, 2014; Meuffels, 1994; Pollmann et al., 2012). However, it is the first time that this is empirically investigated. That a benchmark scale, once constructed, can be used for more than one task is a huge benefit for educational practice because the development of a benchmark rating scale is quite complex. It is important that the benchmarks are good representatives for the different performance levels, and this selection process takes time and expertise (for an overview of how to construct a benchmark rating scale, see for example Feenstra, 2014 or Pollmann et al., 2012). Moreover, when the same benchmark rating scale can be used to rate texts from different writing tasks within the same genre, it is possible to monitor and compare students' writing performance from task to task (cf. Tillema et al., 2012). The results from the present study are therefore promising for large-scale implementation of benchmark scales in the educational system.

The present study demonstrated that raters are able to rate text quality in a reliable manner, even if they are novices and not trained. Compared to the findings of Diederich et al. (1961) in which the average correlations between raters were only .30 for novices and .41 for experienced raters, inter-rater reliabilities in the present study were quite high, showing correlations between two raters of .65 and higher. Whereas the raters in the study of Diederich et al. (1961) received no instructions on how to rate text quality at all, the raters in the present study were at least instructed on what features they had to pay attention while rating the quality of the text. This might have increased inter-rater reliability. This is in line with the results from Barkaoui (2011), who showed that the rating scale type (holistic versus analytic) had larger effects on rater processes than rater experience. Rater agreement can further be enhanced when different points on a rating scale are explained, practiced, and discussed in a training session. This will enhance a common interpretation of how to interpret and apply the levels of a rating scale. However, even when raters are trained carefully, there will still be some degree of disagreement between raters (Cooper, 1977; Weigle, 2002). Moreover, training sessions are criticized for their focus on rater reliability at the expense of validity (cf. Huot, 1990b, Lumley & McNamara, 1995). It is therefore promising that, in the present study, raters were able to rate text quality in a reliable manner, even without training sessions. But, for a satisfactory level of reliability, at least two raters have to be involved in the measurement of writing performance.

Benchmark scales can be used for educational purposes when the different performance levels are standardized (cf. Jonsson & Svingby, 2007). As writing scores can at best provide information at interval level, they do not indicate whether a student's writing performance meets the required standard. For instance, an average score on a benchmark scale does not necessarily indicate that a student is an average writer. For decisions about students' writing proficiency, standards must be defined and additional analysis must be conducted to determine which scores match with what level of quality (Mullis, 1984). Therefore, further research on how to standardize and normalize benchmark rating scales is needed.

Although benchmark ratings seem to be a promising approach to measure students' writing performance, they do not totally solve the rater problem. Cooper and Odell (1977, p. xii) concluded, "since writing is an expressive human activity, we believe the best response to it is a receptive, sympathetic human response". The present study contributes to making this human response more reliable and more valid by supporting raters with benchmarks that illustrate the different performance levels. By doing so, a benchmark rating procedure leads to text quality scores that are a good indication of students' writing performance.



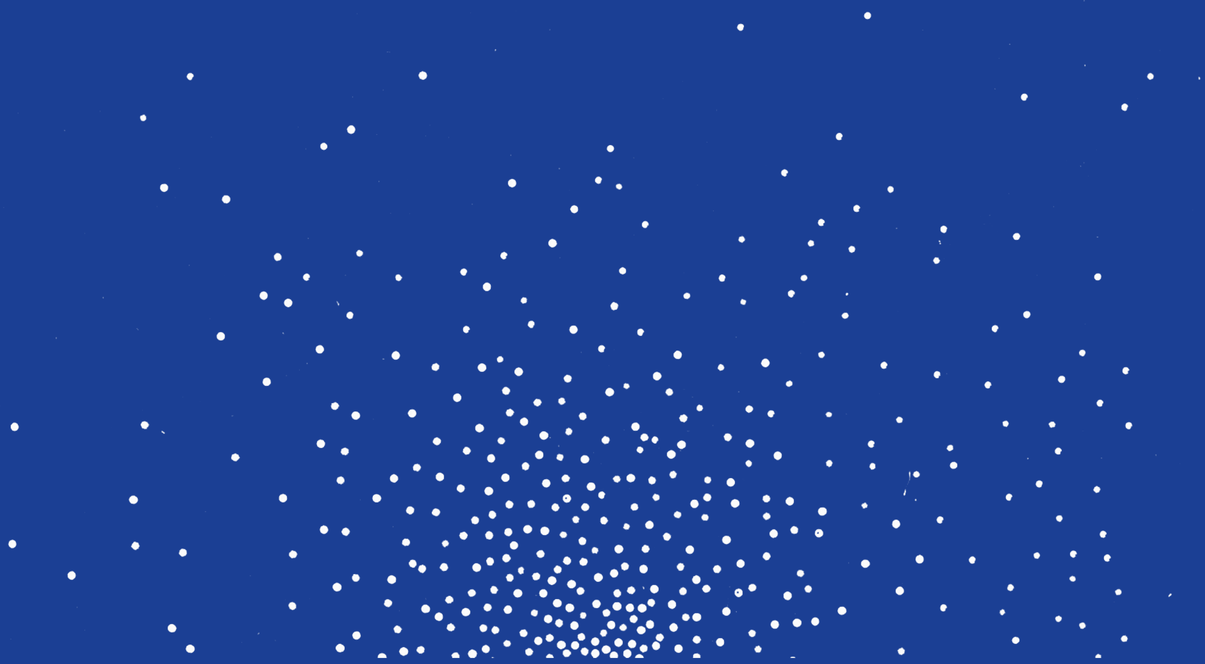
Beste meester ~~Han~~ Han.
Ik wil een uitje plannen naar het pretpark
Een beetje frisse lucht.
Dus volgens mogen' we naar toverland?
Mischin kunnen we wat ~~leren~~ van de
attracties leren?

gr Ties

Chapter 5

EFFECTS OF A STRATEGY-FOCUSED INSTRUCTIONAL PROGRAM ON THE WRITING QUALITY OF UPPER ELEMENTARY STUDENTS

This chapter is based on: Bouwer, R., Koster, M., & Van den Bergh, H. (2016). Effects of a strategy-focused instructional program on the writing quality of upper elementary students. Manuscript is revised and resubmitted for publication.
.



In this intervention study, we tested the effects of Tekster, a comprehensive strategy-focused instructional program for the teaching of writing, incorporating several research-supported practices. The intervention was implemented by teachers ($N=76$) over an eight-week period in 60 4-6th grade classrooms, using a switching replication design with two groups and three measurement occasions. Students ($N=688$) and teachers ($N=31$) in Group 1 worked with the intervention program in the first period of eight weeks, between the first and second measurement occasion, students ($N=732$) and teachers ($N=45$) in Group 2 between the second and third measurement occasion. Students' writing quality improved significantly across all grades in both groups and across all proficiency levels (ES=0.40). This improvement was maintained in Group 1 two months after the intervention. Because writing quality was measured with nine writing tasks in three different genres, findings are generalizable not only over students and classes, but also over tasks. The present study demonstrates that a strategy-focused teaching program, such as Tekster, is a promising approach to improve elementary students' writing.

INTRODUCTION

Despite the fact that writing plays an important role in academic and career success, research has shown that large numbers of students, from many different countries, fail to develop important writing skills (cf. Department for Education, 2012; Salahu-Din, Persky, & Miller, 2008). For example, a recent national assessment in the Netherlands revealed that most elementary-aged students were unable to write texts that conveyed a single, simple message to the reader, and that students writing skills improved negligibly from fourth to sixth grade (Kuhlemeier, Van Til, Hemker, Klijn, & Feenstra, 2013). Further, the Dutch Inspectorate of Education established that only in one third of the schools the teaching of writing is of sufficient quality (Henkens, 2010). Thus, an improvement in writing education at elementary school in the Netherlands is required. For this purpose we developed an instructional program that incorporates several research-supported practices for teaching writing in the upper elementary grades (grade 4 to 6) in which both the focus of instruction (what do we teach) and the mode of instruction (how do we teach it) are addressed. In this study the effectiveness of this writing program is tested.

Focus of instruction

The major problem that developing writers face is cognitive overload. During writing a writer has to perform several resource-demanding cognitive activities simultaneously, such as activate prior knowledge, generate content, planning, formulating, and revising, whilst taking into account the communicative goal of the text and the intended audience (Fayol, 1999). Additionally, the amount of attention required for foundational skills (e.g., handwriting, spelling, sentence and paragraph construction) needs to be considered. This is particularly relevant with developing writers, as they often lack automaticity in these areas (McCutchen, 2011). To circumvent the cognitive overload, developing writers predominantly employ a so-called knowledge-telling strategy in which their thinking about what to write occurs concurrently with production and is typically exclusively content focused (Bereiter, Burtis, & Scardamalia, 1988). However, in this approach text production is restrained by the storage and retrieval capacity of short-term memory (Miller, 1956), resulting in texts that are not sufficiently adapted to the communicative goal and the intended audience (Berninger, Yates, Cartwright, Rutberg, Remy, & Abbott, 1992; McCutchen, 1996). Thus, to improve the writing performance of developing writers, instruction should be aimed at providing them skills and knowledge to manage the cognitive overload during writing.

Strategy instruction. An effective way to manage the cognitive overload during writing is to reduce the number of cognitive processes that are active at the same time, e.g., by teaching students to adopt writing strategies (Kellogg, 1988, 2008). For example, when students are taught to plan during the prewriting phase, they can focus on non-planning processes during writing. Explicit strategy instruction, such as teaching students planning, revising and/or editing strategies, has been the focus of a substantial amount of writing intervention research across a broad range of grades. The strategies under examination were quite diverse: varying from teaching students generic strategies for brainstorming (e.g., Troia & Graham, 2002) or revising (e.g., Fitzgerald & Markham, 1987) to teaching students strategies for specific types of writing tasks, such as writing a story (e.g., Brunstein & Glaser, 2011) or a persuasive essay (e.g., Wong, Hoskyn, Jai, Ellis, & Watson, 2008). Nevertheless, despite these considerable differences, studies involving explicit strategy instruction invariably yield large effect sizes (ESs ranging from 0.82 to 1.15), as is demonstrated in several meta-analyses (Graham, 2006; Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007; Hillocks, 1984; Koster, Tribushinina, De Jong, & Van den Bergh, 2015).

Self-regulation. When strategy instruction is combined with the teaching of self-regulatory skills, effect sizes for strategy instruction are even higher (ES = 1.17; Graham et al., 2012). Self-regulation is “the process whereby individuals activate and sustain behaviors, cognitions, and affect, which are systematically oriented toward the attainment of goals” (Schunk, 2012, p. 123). Essential self-regulatory skills in writing are setting goals for writing, communicative goals as well as process and progress goals, and subsequently monitoring the progress towards these goals (Flower & Hayes, 1981). The most prominent and well-researched approach combining strategy instruction and the teaching of self-regulation skills is the Self-Regulated Strategy Development (SRSD) approach (Harris, Graham, Mason, & Saddler, 2002). In SRSD students are taught strategies for planning, writing, revising and editing, and they are supported in the development of the self-regulation procedures needed to monitor and manage their writing. This instructional approach has been implemented in small groups and whole classrooms, and has invariably proven to be very effective in improving students’ writing performance (Harris et al., 2002).

Students’ self-regulation is positively affected by the attainment of specific goals, which, in turn, will lead to the enhancement of self-efficacy in writing (Latham & Locke, 1991; Schunk, 1990). Students benefit the most from difficult but obtainable goals that make clear what they are to accomplish with their writing before they start, such as establishing the communicative purpose of the assigned writing task and/or criteria for the final product (Schunk, 1990). Previous research has demonstrated that students’ planning, writing, and revising in persuasive writing tasks is promoted by assigning them concrete goals for improving the content of their texts and making them aware of their potential audience (Ferretti, Lewis, & Andrews-Weckerly, 2009; Ferretti, MacArthur, & Dowdy, 2000; Graham, MacArthur, & Schwartz, 1995; Midgette, Haria, & MacArthur, 2008). It is consistently found that specific short-term goals enhance performance better than goals that are general (‘do your best’), long-term goals, or goals that are perceived as too easy or too difficult (cf. Latham & Locke, 1991).

Text structure instruction. Ultimately, to become proficient writers, students have to be able to set their own goals for writing. To do this effectively, students need to know what (communicative) goals should be set for which type of text and how you write a text meeting these goals. For this, students need to have knowledge about text structures and criteria for a good text. For instance, Schoonen and De Glopper (1996) found that proficient writers possess more declarative knowledge about writing and focus more on text structure and organization, whereas poor writers focus on foundational aspects, such as mechanics and layout. The effect of explicit text structure instruction, in which the elements and organization of text types are specifically taught, has been extensively examined in the elementary grades, in different genres: narrative (Fitzgerald & Teasley, 1986; Gordon & Braun, 1986), persuasive (Crowhurst, 1990, 1991; Scardamalia & Paris, 1985), and informative (Bean & Steenwyk, 1984; Raphael & Kirschner, 1985). Meta-analyses (Graham et al., 2012; Koster et al., 2015) show that the overall effect of text structure instruction was positive (ESs 0.59 and 0.76 respectively), even though there was considerable variation between studies in text types.

Mode of instruction

In the way writing education is currently organized, learning to write and task execution are inextricably linked. Young writers face the double challenge of having to produce a text and, at the same time, learning from this activity how to write as well. For developing writers text production is already so demanding, that there is little attentional capacity left for learning from task execution (Rijlaarsdam & Couzijn, 2000). To optimize the way writing is taught, we need to reconsider the organization of writing instruction.

Observational learning. Learning can be effectively separated from task performance by observational learning (Zimmerman & Risemberg, 1997). Observing others performing complex and unfamiliar tasks is less demanding for working memory than actually performing these tasks, in particular when learning complex cognitive skills such as writing (Rijlaarsdam, 2005). Observational learning was first described and studied in the social cognitive learning theory by Bandura (1986). In this theory the idea is emphasized that through observation individuals gain insight into the usefulness and consequences of the behavior that is being modeled. Behavior that is evaluated positively and considered useful will be retained (Schunk, 2012). Observational learning can be applied in teaching writing in two ways: through different types of modeling before and during writing, and through reader feedback during and after writing.

Teacher modeling. In writing instruction observational learning is frequently applied by means of teacher modeling. Modeling involves explaining, demonstrating and verbalizing one's thoughts and actions with the aim to elicit behavioral change in an observer (Schunk, 2012). This kind of modeling prepares students for the forthcoming writing task in the initial phase of the writing process before task execution. Various studies have demonstrated the effectiveness of teacher modeling as an instructional mode to teach writing strategies (cf. Fidalgo, Torrance, Rijlaarsdam, Van den Bergh, & Lourdes Álvarez, 2015; Graham, Harris, & Mason, 2005).

Mastery versus coping models. Models can show either mastery or coping behavior. Mastery models show a flawless performance, whereas coping models initially display exemplary deficiencies of observers but overcome these difficulties and gradually improve their performance (Schunk, 1987; Zimmerman & Kitsantas, 2002). In a study on revision skills, Zimmerman and Kitsantas (2002) found that a coping model raised students' self-efficacy and enhanced their performance more effectively than a mastery model. Research has shown that observing coping models is especially beneficial for weaker students: this may be due to the explicit modeling of strategies to overcome difficulties, or it might be that, due to perceived similarity to the model, students believe that they are also able to improve their performance (Schunk, 1987).

Peer modeling. When peers are used as models instead of teachers, perceived model-observer similarity is even higher, as peers bear close resemblance in development to the observer. Research suggests that peer modeling may be more effective in enhancing self-efficacy and motivation in weaker students, especially when coping models are used (Schunk, 1987). Peer modeling has been investigated in several studies. Raedts, Rijlaarsdam, Van Waes, and Daems (2007) found that observing video-based peer models led to better organized texts, and had a positive effect on students' self-perception of their writing performance. Couzijn (1999) demonstrated that observing peer models led to large learning effects on argumentative text-writing. Further, Van Steendam, Rijlaarsdam, Van den Bergh, and Sercu (2014) found that both proficient and less proficient learners profit from modeling in a collaborative revising task. Braaksma (2002) and Braaksma, Rijlaarsdam, Van den Bergh, and Van Hout-Wolters (2010) found positive effects of observing peer models on both students' writing performance and writing processes. After having observed peer models, students displayed a more mature writing process, and showed a changing pattern of task execution compared to students in the control condition. Furthermore, the findings of Braaksma (2002) support the model-observer similarity hypothesis: Weaker students performed better after focusing on a weak model, whereas good students improved more after focusing on a good peer model. Observing mastery models may be especially beneficial for good students, because they set positive standards for performance (cf. Zimmerman & Kitsantas, 2002). It should be noted, however, that all mentioned studies investigating peer modeling in writing were conducted with secondary or college students, none of these studies involved elementary students.

Reader reaction. Whereas teacher and peer modeling primarily focus on teaching students aspects of the writing process, observational learning can also be applied to provide students feedback on the communicative effectiveness of their written product. Contrary to oral communication, a writer receives no direct cues or feedback from a communicative partner on the communicative adequacy of the text produced during writing, as the writer and reader usually are separated from each other in time and space (Rijlaarsdam et al., 2008). Beginning writers are often unaware of the communicative deficiencies in their own writing. With observational learning, an opportunity can be created to confront writers with the direct impact of their text, and the deficiencies therein, on a reader. Observing genuine readers and discussing readers' experiences provide students with valuable information on the readers' needs and whether they succeeded in fulfilling these needs (Couzijn & Rijlaarsdam, 2004; Schriver, 1992). Several researchers (Couzijn, 1995; Couzijn & Rijlaarsdam, 2004; Holliway & McCutchen, 2004; Rijlaarsdam, Couzijn, Janssen, Braaksma, & Kieft, 2006) have shown that students' writing improved when they experience the effect their text has on a reader. Meta-analyses indicated that feedback and peer interaction when writing are highly effective: Peer interaction when writing was associated with effect sizes of 0.59 (Koster et al., 2015) and 0.89 (Graham et al., 2012), and feedback with effect sizes of 0.80 (Graham et al., 2012) and 0.88 (Koster et al., 2015).

Gradual release of responsibility. The improvement of writing performance cannot be accomplished with observational learning alone. At one point in time, after having observed models, students have to start writing themselves. Where initial modeling by the teacher or by peers facilitates students to commence writing, they still have to progress through all the stages of the writing process to successfully complete the writing task. Thus, there still is a gap to be bridged: from modeling by a more knowledgeable other to students' independent practice. This can be achieved through gradual release of responsibility (Pearson & Gallagher, 1983), by which the cognitive load gradually shifts from modeling to guided practice and, finally, independent performance. The gradual release of responsibility model builds on Vygotsky's sociocultural theory and the concept of the zone of the proximal development (Vygotsky, 1980). Vygotsky (1980) defines the zone of proximal development as the area between the student's level of independent performance and the student's level of potential development as determined by assisted performance. A teacher can facilitate the learner's progression from assisted to independent performance through scaffolding (Wood, Bruner, & Ross, 1976). In scaffolding the teacher controls the elements of the task that are initially beyond the student's capacity, thus permitting the student to concentrate upon the elements that are within his range of competence (Wood et al., 1976). The amount of teacher assistance can gradually be decreased as the learner progresses. For successful scaffolding it is essential that teachers help students to develop strategies that can be transferred to new tasks and situations (Bodrova & Leong, 1998).

Intervention programs that use the gradual release of responsibility and scaffolding, have been successful in improving students' writing performance (cf. Graham et al., 2005; Graham et al., 1995). In these interventions, in addition to modeling, also explicit instruction is used to activate background knowledge, for instance to identify criteria for a good text. Further, explicit instruction by the teacher is applied to make students aware of the purpose and benefits of the strategies that are taught. Research has shown that students are more likely to use a strategy in new writing situations if they are aware that its use leads to improved writing. For developing writers (i.e., grade 5 & 6) transfer of strategies to other tasks or domains can be enhanced by more comprehensive and explicit instructions regarding how and when the strategy can best be applied (O'Sullivan & Pressley, 1984).

Aim of the study

The main purpose of this study was to test the effectiveness of a comprehensive teaching program, Tekster [Texter] (Koster, Bouwer, & Van den Bergh, 2014a, 2014b, 2014c), in which several research-supported instructional practices are combined into one general overall approach for writing to improve the writing quality of students in the upper grades of general elementary education in the Netherlands. To reduce students' cognitive overload during writing the focus of instruction (what do we teach), as well as the mode of instruction (how do we teach it) are adapted. In Tekster the main focus of instruction is to teach students a strategy for writing, supplemented with the teaching of self-regulatory skills, and explicit instruction in text structure. The predominant mode of instruction is observational learning, complemented with instruction and guided practice with extensive scaffolding, following the gradual release of responsibility model (Wood et al., 1976). The approach of Tekster bears close resemblance to the approach that is applied in SRSD (Harris et al., 2002) and CSRI (Cognitive Self-Regulation Instruction, see Fidalgo et al., 2015). As it is the aim to improve students' overall writing performance, we teach students a general writing strategy and address the genre-specific features through text structure instruction. During the intervention students practice in writing several different types of texts, such as a story, a recipe, a letter, and a brochure. Consequently, to be able to make claims about any improvement in students' overall writing performance, writing proficiency has to be measured with multiple writing tasks and multiple text types (Bouwer, Béguin, Sanders, & Van den Bergh, 2015). So, in this study several text types were used to measure writing quality: descriptive texts, narratives, and persuasive letters.

In the present study we investigated whether Tekster improved the writing performance of students in the upper grades (grade 4-6) of elementary school, in a general educational setting, and whether this effect differed between grades. We also examined whether the effect of the intervention was maintained over time. Further, we investigated if there were differences in the effect of the intervention between male and female students, as girls are generally better writers than boys (cf. Berninger & Fuller, 1992; Pajares & Valiante, 1999). Finally, we investigated whether the effect of the program depends on the proficiency of the writer.

METHOD

Sample

In total, 76 teachers of 60 classes volunteered to participate in the study. The majority of teachers were female (82%), which does not differ from the percentage female teachers in the population (85%; $\chi^2(1) = 0.04$, $p = .84$; Inspectorate of Education, 2012). All teachers were qualified and experienced elementary teachers. They were from 27 schools spread all over the Netherlands: 11 were located in the northern part, 9 in the middle region and 7 in the southern region. Schools varied in their identity: 16 schools (60%) were grounded in a religious denomination (11 Catholic, 2 Protestant, 2 Reformed, 1 Islamic) and 40% were public schools. The sample of schools reflects the population in which two-thirds of the schools are denominational schools and one-third of schools is public (Ministry of Education, Culture and Science, 2015). There were 10 schools of which only one class participated, 6 schools with two classes, 8 schools with 3 classes, 1 school with 4 classes and 2 schools with 5 classes. Of these classes, 20 were fourth grade classes, 13 fifth grade classes, 16 sixth grade classes and 11 multigrade classes combining two or three grades.

In total, 1420 students participated in the study: 477 fourth grade students (mean age = 9.40, *SD* = 0.62), 454 fifth grade students (mean age = 10.40, *SD* = 0.61), and 489 sixth grade students (mean age = 11.50, *SD* = 0.64). In our sample the average number of students per class was 23.6 students (*SD* = 5.6), which does not deviate from the population (23.3 students, see Central Office for Statistics, 2015). The average percentage of female students per class in the sample, which was 50%, also reflects the average in the population (Central Office for Statistics, 2015). Specific information on individual students' SES or special needs was not available. In general, an average Dutch classroom consists of 20 to 25% of special needs students (Koopman, Ledoux, Karssen, Van der Meijden, & Petit, 2015). These are students needing additional individual attention and/or care, such as students with learning disabilities or gifted students.

Attrition. Initially 64 classes participated in this study, however, due to time constraints in their schedules, four teachers from three schools (one school from Group 1, and two schools from Group 2) reported at the start of the intervention not being able to work with the program in their classroom after all. A small number of individual students dropped out during the study because they changed schools. In total, 37 students (2.6%) missed one of the two posttests and only 17 of them (1.2%) missed both posttests.

Design of the study

To analyze whether the writing program improved students' writing skills, a design with switching replications (Shadish, Cook & Campbell, 2002) was used, including two groups and three measurement occasions, see Table 5.1. In the first phase of the study, from pretest to the first posttest, teachers and students in Group 1 worked with the intervention program for eight weeks, two lessons a week, instead of their regular program for writing. Group 2 served as a control group during this period in which teachers and students engaged in their existing writing activities and routines. During the second phase of eight weeks, between the second and third measurement occasion, the intervention switched between groups, such that the teachers and students in Group 2 started to work with the writing program, while those in Group 1 returned to their regular writing activities. The third measurement occasion served as a posttest for students in Group 2, as well as a delayed posttest for students in Group 1, with which we were able to measure retention.

Table 5.1 Switching replication design with two groups and three measurement occasions

	M1	Phase 1	M2	Phase 2	M3
	Tasks	(8 weeks)	Tasks	(8 weeks)	Tasks
Group 1	a, b, c	Intervention program	d, e, f	Regular program	g, h, i
Group 2		Regular program		Intervention program	

A switching replication design is superior to a regular pre-post (quasi-) experimental design as the intervention is implemented in both groups but in different time intervals. It is not only a more ethical design, as all students eventually benefit from the intervention, but it also allows to test for internal validity. If the intervention is equally effective in both groups, the effect does not depend on (characteristics of) a particular group. If the effect of the intervention is not equally effective in both groups, internal validity might be threatened. Moreover, because the intervention is replicated in two groups, it yields important information about the reproducibility and generalizability of the results, for which socials scientists recently called more attention (Open Science Collaboration, 2015). Furthermore, this design provides information on the maintenance of the effects of the intervention, as it includes a delayed posttest for students in Group 1.

Assignment of schools to groups. The school holiday calendar determined which schools were assigned to Group 1 or Group 2. Specifically, schools located in the northern region were assigned to Group 1 and those located in the southern region were assigned to Group 2. Schools from the middle region were randomly assigned to Group 1 and 2. Group 1 contained 14 schools with 31 teachers (84% female) and 29 classes, Group 2 contained 13 schools with 45 teachers (80% female) and 31 classes. Group 1 received the intervention in the first period of eight weeks and Group 2 received the intervention in the second period of eight weeks. Student information per group is presented in Table 5.2. The number of students per grade was similar for both groups ($\chi^2(2) = 2.67$, $p = .26$), and there were no significant differences in the percentage of boys/girls between the two groups ($\chi^2(1) = 2.21$, $p = .14$) nor in their age ($t(1414) = -1.31$, $p = 0.19$).

Table 5.2 Students' characteristics per group and per grade

	Group 1			Group 2		
Grade	*N*	*% female*	Mean age (*SD*)	*N*	*% female*	Mean age (*SD*)
4	245	47%	9.41 (0.58)	232	54%	9.39 (0.65)
5	217	51%	10.39 (0.63)	237	54%	10.42 (0.59)
6	226	46%	11.50 (0.67)	263	48%	11.50 (0.62)
Total	688	48%	10.41 (1.07)	732	52%	10.48 (1.07)

Writing instruction

Existing instruction. In the present study, the intervention program is compared to the existing classroom practice in writing education. In the Netherlands, writing education traditionally is part of the Dutch language teaching curriculum. A report of the Dutch Inspectorate of Education (Henkens, 2010) demonstrated that from the 8 hours per week reserved for language teaching, on average 45 minutes are spent on writing. These writing lessons are primarily product-focused: Students receive hardly any support during the writing process, nor are they supported on how to approach writing tasks. Further, the report showed that in the majority of schools the writing performance of students is not monitored, and students are seldom given feedback on their performance. These findings were corroborated by a recent study on classroom practice in teaching writing in the Netherlands with 51 elementary school teachers (Rietdijk, Van Weijen, Janssen, Van den Bergh, & Rijlaarsdam, 2015). In this study the majority of teachers (94%) spend less than one hour a week on teaching writing. Further, this study shows that teachers pay somewhat more attention to elements of the writing process than was reported by the Inspectorate in 2010: a majority of teachers indicate that they promote prewriting activities, such as the generation of ideas, and half of the teachers asks students to revise their texts. However, during an average lesson students work mostly individually; only one-third of the lesson time is devoted to plenary instruction. Teachers hardly offer any individual support for students, nor do they apply modeling in their writing instruction. Only a minority of teachers provide feedback on the communicative effectiveness of the text (Rietdijk et al., 2015).

Intervention: Tekster. The intervention consisted of a teaching program, Tekster, which included three lesson series of 16 lessons, one for each grade level, compiled in a workbook, accompanied by a manual and a introductory training session for participating teachers (Koster et al., 2014a, 2014b, 2014c). In the Tekster-program several effective practices are combined to address the focus as well as the mode of writing instruction. Table 5.3 gives a general description of the design principles of the program and how these principles were operationalized in learning and teaching activities in the program (see Rijlaarsdam, Janssen, Rietdijk, & Van Weijen, in press). In particular it shows how we combined and

Table 5.3 Design principles for focus and mode of instruction of the teaching program Tekster and their translation into learning and teaching activities

Design principles		**Teaching program**	
Focus of instruction	**Mode of instruction**	**Learning activities**	**Teaching activities**
1. Process related: Writing strategies *General approach for writing tasks, based on phases of the writing process: generate content, organize, formulate, reread, evaluate, revise*	a. Observational learning	Observe/discuss/compare model(s), (teacher or peer) applying the writing strategy in different stages of the writing process	Modeling strategy use (thinking aloud while performing (part of) the writing task)
	b. Explicit instruction	Listen actively, retrieve relevant background knowledge from memory, take notes	Explain the components of the strategy, make students aware of the purpose and benefits of using writing strategy, activate student's background knowledge
	c. (Guided) practice	Apply the steps of the strategy to writing tasks: Authentic tasks with clear communicative goal and intended audience in various genres	Provide help when needed through scaffolding and process feedback

▼

Design principles		Teaching program	
Focus of instruction	Mode of instruction	Learning activities	Teaching activities
2. Product related: Text structure *Criteria for written product, depending on communicative goal and intended audience*	a. Observational learning	Before writing: Observe/discuss/compare model(s), (teacher or peer) talking about criteria for and conventions of various text types, compare and discuss model texts of the same text type to derive criteria and conventions for a good text	Before writing: Model the relevant aspects of the text type, provide model texts or show video clips of peer modeling
		After writing: Evaluate peer/own text on the basis of the previously discussed criteria and give feedback (reader reaction), observe reader reaction, observe model revising on the basis of feedback	After writing: Evaluate students' texts on the basis of previously discussed criteria, give feedback (reader reaction), model how feedback can be used in revision to improve the text
	b. Explicit instruction	Listen actively, take notes	Explain why and how the criteria and conventions should be used, discuss important criteria and conventions on the basis of model texts
	c. (Guided) practice	Apply the discussed criteria to writing tasks: Authentic tasks with clear communicative goal and intended audience in various genres After writing: Give feedback /assess own text according previously discussed criteria	Provide help when needed through scaffolding and product feedback

Design principles		Teaching program	
Focus of instruction	Mode of instruction	Learning activities	Teaching activities
3. Writer related: Self-regulation skills *Writer's monitoring and regulating of own progress in relation to communicative goals*	a. Observational learning	Observe/discuss/compare model(s), (teacher or peer) setting goals and monitoring progress in relation to goals during the writing process, observe/discuss/compare effect of self-regulation on the written product.	Model self-regulation during writing, by thinking aloud during performing writing task
	b. Explicit instruction	Listen actively, take notes	Explain why it is important to set communicative goals for writing in advance, explain the differences between various communicative goals, when and how during the writing process progress towards the communicative goal can best be monitored
	c. (Guided) practice	Set communicative goal before writing, monitor progress towards this goal during writing, regulate own writing process and adapt if necessary, evaluate written product in relation to communicative goal, revise if necessary.	Provide help when needed through scaffolding, and self-regulation feedback

translated research-supported practices for mode and focus of instruction into learning activities for students and teaching activities for teachers.

General lesson format. The lessons of the program are designed following a more or less fixed format, see Table 5.4 for the general Tekster-lesson format. This table demonstrates how the components of Table 5.3 are integrated in the actual lessons. As can be seen in Table 5.4, the core of the lessons is the overall writing strategy. To support students in applying the strategy, they are taught a mnemonic representing the steps of the writing process: VOS (fox) for grade 4, DODO (dodo) for grade 5, and EKSTER (magpie) for grade 6. The letters of the acronyms represent the steps in the writing process as follows: VOS (fox) for Verzinnen (generate content), Ordenen (organize), Schrijven (write); DODO (dodo) for Denken (think), Ordenen (organize), Doen (do), Overlezen (read); Ekster (magpie) for Eerst nadenken (think first), Kiezen & ordenen (choose & organize), Schrijven (write), Teruglezen (reread), Evalueren (evaluate), Reviseren (revise). These animals are the common theme throughout the lessons, with small icons of the animals serving as a visual support. The duration of the average Tekster-lesson is between 45 and 60 minutes. A sample lesson is included in Appendix M.

Table 5.4 General lesson format of Tekster

Lesson phase	Activities
1	Goal of the lesson is explicitly stated (3b)
2	Plenary introduction in which specific characteristics of text type are addressed (through modeling (2a), comparing model texts (2a), or explicit teacher instruction (2b))
3	Introduction of authentic writing assignment in which communicative goal and intended audience are explicated (3b)
4	Acronym for the strategy is explicitly named (1b)
5	First step of the strategy: generate content in keywords (gradual release of responsibility from 1a to 1c, 3a to 3c)
6	Second step of the strategy: organize content in keywords (gradual release of responsibility from 1a to 1c, 3a to 3c)
7	Third step of the strategy: write the text using organized content (1c, 2c, 3c)
8	Fourth step of the strategy (grade 5 & 6): students read each other's text or their own text (2a)
9	Fifth step of the strategy (grade 5 & 6): students evaluate the text by answering evaluative questions and/or giving feedback (2a)
10	Sixth step of the strategy (grade 6): students revise (parts of) their text on the basis of the feedback they received (3c)

Note. Numbers between parentheses refer to focus and mode of instruction as displayed in Table 5.3.

Content of the writing lessons. In the first lesson the acronym-animal introduces itself in a story in which students also practice the steps of the strategy for the first time. In the following lessons students learn to apply the writing strategy to various types of texts, which are authentic writing tasks with various communicative goals and audiences. For instance, in each grade they learn to write descriptive texts (e.g., a personal ad or self-portrait), narrative texts (e.g., a story for kindergarten or newspaper article), persuasive texts (e.g., a nomination email for a television program or a flyer for recruiting new members for a club), instructive texts (e.g., a recipe, rules for a game) and personal communication (e.g., a holiday postcard or an invitation for a party). The writing tasks of the program comply with the goals set for the end of elementary school by the Dutch Ministry of Education. As a target goal for the end of elementary school the Ministry proposes that "students are able to write coherent texts, with a simple linear structure on various familiar topics; the text includes an introduction, body, and ending" (Expert Group Learning Trajectories, 2009, p.15).

The level of difficulty ascended through the grades as follows: In grade 4, writing tasks were used in which the intended audience was in close proximity of the student, such as classmates, friends, or (grand-) parents. In grade 5, this was expanded to people with whom students have a more distal relationship, but are still familiar to them, such as their teacher, relatives, or neighbours. In grade 6, students also had to write texts that were intended for unfamiliar people, such as the editor of a newspaper, or the managing director of a company.

Development of the lessons. The content of the lessons was developed in close collaboration with sixteen elementary school teachers, divided into three teacher-design-teams, who had monthly meetings over a period of six months. After an introduction to the didactical principles of the program, two teams were given the general lesson format with the instruction to develop writing assignments suitable for this approach. The third team made the peer modeling video clips. Requirements for the writing assignments for the lessons were that they had a clear communicative goal and audience, and that the topics would match the students' interest and developmental level. The teachers piloted the writing assignments with their own students, and received feedback from their team members and the researchers during the monthly meetings. Subsequently, the program was tested in a pilot study (Koster, Bouwer, & Van den Bergh, 2016a).

Teacher instruction. To support teachers in the implementation of the program we provided one introductory four hour session training session prior to the start of the intervention, during which teachers were trained in small groups (max. 12 teachers). In this training session, all teachers received a teacher manual to facilitate the teaching of Tekster-lessons. The manual consisted of two parts: a general introduction and detailed lesson plans for each lesson. In the general introduction the goal and approach of the program were explicated as well as the general lesson format. Further, the importance of feedback for learning to write was explained and suggestions were given on how to provide effective feedback. Next, observational learning and modeling were explained, and practical information was given concerning the organization of the experiment. The detailed lesson plans provided an overview of the instruction and activities of the lesson with a time planning for each phase of the lesson. These lesson plans described the activities the teacher was expected to execute during the lesson and provided suggestions when to use modeling during instruction. The manual also included a dvd with movie clips of peer modeling to use during instruction, and videos with examples of teacher modeling for different phases of the lesson.

The teacher manual was used as a guidance during the training session, to brief teachers on how to work with the program. The researchers informed the teachers about the theoretical background, as explained in the manual, and showed and discussed an example video of teacher modeling. Next, teachers discussed examples of feedback. For effective feedback teachers have to adjust their comments to students' needs, which requires that teachers are able to assess the quality of students' texts and adapt their feedback accordingly (Bouwer, Koster, & Van den Bergh, 2016b). Therefore teachers were trained in how to evaluate text quality using benchmark texts, and how to use this information to provide effective feedback. Subsequently they practiced giving feedback on example texts, and reflected on the quality of their feedback in subgroups. At the end of the training session teachers were instructed to read the information in the manual carefully and watch the videos before the start of the program as a preparation for teaching the lessons.

The teacher manual also contained a logbook after each lesson, including questions concerning the practical implementation of the intervention in the classroom, which provide valuable insight in the social validity of the intervention. Teachers were asked to fill out the following information: preparation time, appreciation score for the lesson as a whole (on a scale from 1 to 10; 1=very low, 10=very high), the estimated level of difficulty for their students (on a 5-point scale: 1=easy, 5=hard), the level of difficulty of executing the lessons (on a 5-point scale: 1=easy, 5=hard), and additional comments (if any).

Intervention fidelity

We included fidelity measures to examine whether teachers implemented the program as intended. In this study, fidelity was operationalized as the number of lessons that were provided by each teacher, as well as the number of completed lessons of each student. We also checked whether teachers adhered to the lesson plans as provided in the teacher manual. We collected and reviewed all students' workbooks after the intervention to determine the number of lessons that were completed by each student. Further, we collected and reviewed teacher entries in logbooks to get more insight into the number of lessons taught and to determine whether teachers adhered to the time schedule provided in the lesson plan. Additionally, in a random sample of two-third of the classrooms we performed classroom observations to determine whether teachers followed the lesson plans.

Workbooks. Of each student's workbook, trained undergraduate students coded the number of completed lessons. If a student had written a text, a lesson was considered as completed. Analysis of this data revealed that the number of completed lessons varied considerably between and within classes. On average, students completed 10 lessons: 8% of the students completed less than four lessons and 53% of the students 10 or more lessons.

Logbooks. Logbooks were included in the teacher manual after every lesson. In the logbooks teachers had to report the duration of the lesson. After the intervention, 75% of the logbooks were returned, and subsequently coded by two trained graduate students. Analysis of the logbook data showed that teachers taught 10 lessons on average. The lesson duration varied from 29 minutes to 58 minutes, the average lesson duration was 43 minutes, which closely approximates the lesson time as prescribed in the manual.

Observations. Our observation instrument, based on the work of Hintze, Volpe, and Shapiro (2002), was designed to map the teacher's activities during the lessons and the implementation of the key elements of the program, i.e., modeling and the use of the writing strategy. Every 20 seconds was tallied whether the teacher was on task or off task. On task if the teacher was executing the actions as specified in the lesson plan for that phase of the lesson, off task if the teacher was involved in other activities than teaching writing, such as fetching a cup of coffee or talking to a colleague. Further, it was tallied whether the on-task-behavior involved plenary activities (instruction or classroom interaction) or interaction with individual students. Additionally, for every lesson phase observers had to register whether the teacher performed teacher modeling and referred to the acronym and/or the steps of the strategy, which are the key elements of the intervention. Each classroom was observed by one trained undergraduate student, there were ten observers in total. To optimize observers' agreement, all observers were trained in advance.

Analysis of the observational data showed that teachers closely adhered to the instructions in the manual and the lesson plans. Teachers were on task on average 92% of the total lesson time. Half of this time (54%) was devoted to plenary instruction, the rest of the time teachers were monitoring progress and providing instruction to individual students. The observation data showed that teacher modeling was applied on average 1.3 times per lesson, and that the acronym (mnemonic for the writing strategy) was used on average 1.4 times per lesson.

Assessment of writing quality

Writing tasks. Because generalization to writing proficiency is not warranted when writing scores are obtained with only one writing task (Bouwer et al., 2015), we assessed students' writing skills at each measurement occasion using three different types of texts: descriptives (tasks a, d, g), narratives (tasks b, e, h) and persuasive letters (tasks c, f, i), see also Table 5.1. For each text type, tasks were as similar as possible, only differing in topic. The tasks were developed by the researchers and were not used in previous studies or in existing educational methods. The quality of the tasks was discussed with experts in the field of writing education. Special attention was devoted to the level of difficulty and whether the topics were interesting and would trigger enough ideas to produce a text of reasonable length. Each task contained the prompt, including an illustration with relevance to the topic, and some space for prewriting which students were free to use. Examples of writing prompts for each text type are provided in Appendix F.

Administration of writing tasks. The writing tasks for the assessment of students' writing proficiency were administered by the teachers, during normal class time. Teachers were briefed to administer the tasks without offering any additional instructions before or during writing, and to let students work individually. They were further instructed to plan the three writing tasks for each measurement occasion within one week, but not on the same day. There was no time constraint explicated for students in which to complete the writing task.

Rating writing quality. As information concerning the writer, e.g., gender, age, ethnicity, might influence the rater, we anonymized all students' texts. However, due to the scope of this study (1420 students each had to write nine texts which resulted in a sample of approximately 12780 written texts) it was not feasible to retype all handwritten texts, and hence, to control for possible presentation effects. Raters were experienced elementary teachers (grade 4 to 6) who assessed global text quality using a continuous rating scale with five benchmarks (Blok & Hoeksma, 1984; Bouwer, Koster, & Van den Bergh, 2016c). The benchmark scale can be considered as an interval scale. The center position on the scale is an average text which is assigned an arbitrary score of 100 points. The other texts on the scale are one (115 points) and two (130 points) standard deviations above average, and one (85 points) and two (70 points) standard deviations below average.

The benchmarks on the scale originated from a preliminary investigation of a randomly selected subsample of the total set of texts written at the first measurement occasion (see Table 5.1). This sample consisted of texts in the three different genres, written by students in all three grades, reflecting the range of writing quality in grade 4 to 6. Five experienced upper-elementary teachers rated the subsample holistically and their scores were averaged. Subsequently, benchmarks were selected based on two criteria. First, the text had to be a good representative of the quality level (-2 SD, -1 SD, 0, +1 SD, +2 SD). Second, there had to be high agreement between raters on the quality of the text (low variance between raters' scores). For each type of text, i.e., descriptives, narratives and persuasive letters, a different benchmark scale was constructed, see Appendix K for an example.

Raters, who were trained in advance on how to use the scale, independently compared students' texts to the benchmarks and scored each text accordingly, blind to experimental condition. Each text was rated by a jury of three raters, using a design of overlapping rater teams. In this design writing products are split randomly into subsamples, equaling the number of raters ($N = 47$). Each rater received 3 subsamples according to a prefixed design. Due to the overlap in samples, the reliability of raters (and juries) can be approximated (Van den Bergh & Eiting, 1989). The average reliability of jury ratings across assignments was high, $\rho = .89$, and varied between assignments from $\rho = .86$ to $\rho = .91$. The final quality score for each text was determined by computing the mean of the scores of the three raters. As scores appeared to be somewhat negatively skewed, i.e., raters tended to score texts of low quality more extremely, raters' scores were normalized for each task (Blom's rank-based normalization formula, see Solomon & Sawilowsky, 2009).

Data analyses

The data in the present study are hierarchically organized; scores are cross-classified with students and tasks, and students are nested within classes. Therefore, the data are analyzed by applying different (cross-classified) multilevel models in which parameters are added systematically to the model. In such models all students, including those with partly missing values, are taken into account.

The effectiveness of the intervention across group and grade is tested with six models. Model 1 is the basic null model in which we only account for random error (S^2_e) and random effects of students (S^2_s), tasks (S^2_t), and classes (S^2_c). That is, writing scores are allowed to vary within and between students, between tasks (including systematic variation due to genre), and between classes. In Model 2 measurement occasion is added as a fixed effect to test whether average scores differ over time. In Model 3 it is tested whether the variances within and between students and between classes differ between the three measurement occasions. In Model 4 group is added as a fixed effect to test whether average scores differ between the two groups. Model 5 tests the main effect of the writing intervention by estimating the interaction between group and measurement occasion (see Table 5.1 for the research design). This model includes the restriction that the effect of the intervention is the same in the two groups. In Model 6 this restriction is removed to test whether the intervention is equally effective in Group 1 and 2, which is in essence a check on the internal validity of the experiment.

The maintenance effect of the intervention is tested by performing a specific contrast analysis of students in Group 1. In this analysis students' writing scores at the posttest are compared to their writing scores at the delayed posttest.

To test the effect whether the intervention is equally effective in different grades, we applied two additional models to the data. In the first model grade is added as a fixed effect to test whether average scores differ between the three grades. In the second model the interaction effect between the intervention (measurement occasion*group) and grade is added to test whether the intervention is equally effective in different grades.

The role of gender on the effectiveness of the intervention is tested by two additional models. In the first model gender is added as a fixed effect to test whether average scores differ between boys and girls. In the second model the interaction effect between the intervention (measurement occasion*group) and gender is added to test whether the intervention is equally effective for boys and girls.

To test the effect whether the intervention is equally effective for students with different abilities, we performed an aptitude treatment interaction analysis. For this analysis, we included students' writing scores at the first measurement occasion as a covariate for students' outcomes at the second measurement occasion (the posttest).

RESULTS

Effect of the intervention

Results of the fit and comparison of the six models are shown in Table 5.5. As can be seen, there was a fixed effect of measurement occasion (Model 2 versus Model 1, $\chi^2(2) = 279.61$, $p < .001$), which indicates that average writing scores were not equal over time. Allowing the variances to differ between measurement occasions significantly improved the model (Model 3 versus Model 2, $\chi^2(12) = 657.61$, $p < .001$). This means that for at least one of the levels (students, tasks, classes and/or random error), the variance was not homogeneous across measurement occasions. There was no main effect of group (Model 4 versus Model 3, $\chi^2(1) = 1.32$, $p = .25$), indicating that average writing scores were the same for students in Group 1 and 2.

Table 5.5 Fit and comparison of nested models

Model	N_{pars}	-2 Log Likelihood	Comparison Models	$\Delta\chi^2$	Δdf	p
1 null	5	88763.76				
2 + measurement occasion (fixed)	7	88484.15	2 vs 1	279.61	2	< .001
3 + measurement occasion (random)	19	87826.54	3 vs 2	657.61	12	< .001
4 + group	20	87825.22	4 vs 3	1.32	1	.25
5 + intervention	21	87800.24	5 vs 4	24.98	1	< .001
6 + intervention * group	22	87800.12	6 vs 5	0.12	1	.73

There was a main effect of the intervention (Model 5 versus Model 4, $\chi^2(1) = 24.98$, $p < .001$), indicated by an interaction between group and measurement occasion. This means that differences in writing scores measured at two occasions (i.e., between the first and second occasion or the second and third occasion) were not the same for students in the intervention condition and the control condition. The effect of the intervention on differences in writing scores appeared to be the same for students in Group 1 and 2 (Model 6 versus Model 5, $\chi^2(1) = 0.12$, $p = .73$). To verify the direction of the interaction effect, two additional contrasts were performed. The results showed that between the first

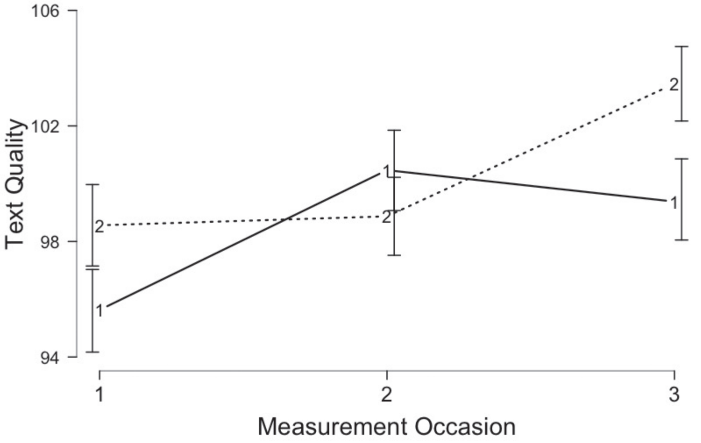
Table 5.6 Students' average writing scores and variances on pre- and posttest measures

	Measurement occasion		
Dimension	1	2	3
Fixed part			
Group 1	95.63 (1.38)	100.36 (1.36)	99.36 (1.34)
Group 2	98.54 (1.41)	98.78 (1.33)	103.51 (1.28)
Random part			
$S^2_{classes}$	53.92 (11.31)	49.79 (10.44)	43.73 (9.40)
S^2_{tasks}	9.20 (1.42)	9.20 (1.42)	9.20 (1.42)
$S^2_{students}$	59.99 (4.33)	54.98 (3.65)	54.31 (3.65)
S^2_{error}	128.48 (3.68)	99.11 (2.83)	92.98 (2.77)

Note. Standard errors are included in parentheses.

and second measurement occasion the scores of students in the control group (Group 2) remained the same ($\chi^2(1) = 0.09$, $p = .76$), whereas the scores of students in the intervention group (Group 1) increased significantly with 4.73 points ($\chi^2(1) = 71.97$, $p < .001$). The magnitude of this effect was estimated by comparing the effect of the intervention to the total amount of variance (Cohen's d). This resulted in an estimated effect size of 0.32, while generalizing over students, teachers, and tasks. Parameter estimates of the final model are summarized in Table 5.6, and a graphical overview of the effect of the intervention in both groups is presented in Figure 5.1. The parameter estimates in Table 5.6 show that the variance within and between students and between classes decreased over time. The decrease in between-class variance indicates that classes

Figure 5.1 The effect of the intervention on text quality averaged over grades. Error bars indicates 95% confidence intervals for the means. Solid lines represent Group 1, which received the intervention between the first and second measurement occasion. Dashed lines represent Group 2, which received the intervention between second and third measurement.

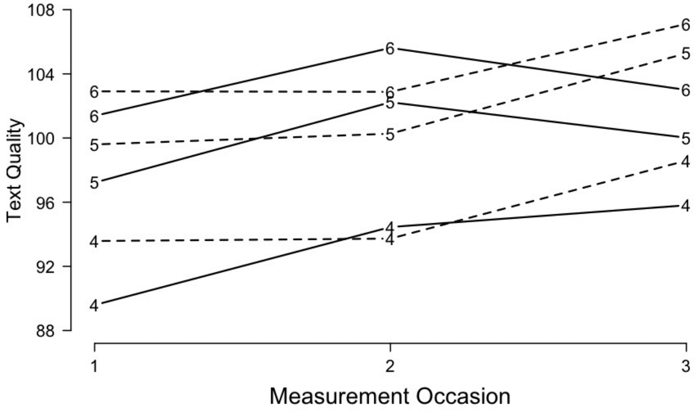


became more homogeneous over time. Smaller within-students variance is related to smaller interaction effects between students and tasks, indicating that students' writing also became more homogeneous.

Inspection of the students' workbooks revealed that the number of lessons performed varied considerably within classes ($M = 10$, $SD = 4$). The estimates of mean scores in Table 5.6 are based on the mean number of completed lessons. To examine whether the effect of the intervention depended on the number of completed lessons, the so-called dosage effect, the number of completed lessons per student was included as a fixed factor in the analyses. Results showed that students' writing scores increased as a function of the number of completed lessons, $\beta = 0.21$ ($SE = 0.09$, $p < .01$), which also affects the effect size. If a student completed the whole program of 16 lessons, her writing quality increases with 5.99 points, which yields an estimated effect size of 0.40.

Grade differences. The results further show a main effect of grade ($\chi^2(2) = 54.40$, $p < .001$). This indicates that average writing scores were different for students in grades 4 to 6. Moreover, the effect of the intervention was not the same for each grade ($\chi^2(2) = 14.21$, $p < .001$). On average, the writing quality of fourth grade students improved by 4.86 points (Cohen's $d = 0.34$), of fifth grade students by 5.00 points (Cohen's $d = 0.35$), and of sixth grade students by 4.23 points (Cohen's $d = 0.30$) after following the intervention program. A graphical overview of the effect of the intervention for grade 4 to 6 is presented in Figure 5.2.

Figure 5.2 The effect of the intervention on text quality and separated by grade (numbers indicate grade). Solid lines represent Group 1, which received the intervention between the first and second measurement occasion. Dashed lines represent Group 2, which received the intervention between second and third measurement.



Maintenance effect. In the present study, the writing proficiency of students in Group 1 was measured twice after the intervention program: once directly after the intervention (posttest) and once after two months (delayed posttest).

Between the posttest and the delayed posttest, students followed the existing writing program. Specific contrast effects indicated that the effect of the intervention maintained over time. Writing scores improved significantly between the pretest (M1) and the delayed posttest (M3; $\chi^2(1) = 23.14$, $p < .001$), but no significant differences in writing scores were found between the posttest (M2) and the delayed posttest (M3; $\chi^2(1) = 2.06$, $p = .15$). This indicates that the effects of the intervention on writing quality persisted over time.

Effect of gender. Results further show a large main effect of gender on the writing scores ($\chi^2(1) = 319.70$, $p < .001$), indicating that girls scored on average 7.62 points higher than boys. Although girls generally outperformed boys on writing tasks, the effect of the intervention was not different, as indicated by a non-significant improvement of the model in which the interaction between group, measurement occasion and gender was allowed ($\chi^2(1) = 0.10$, $p = .75$).

Aptitude treatment interaction

In a separate analysis was investigated whether the effect of the intervention depended on students' writing proficiency. For this analysis the regression of the scores of the first measurement occasion on the second measurement occasion were estimated per group. In Group 1, with an intervention in the first eight weeks, the regression coefficient equaled 0.60 ($SE = 0.02$), whereas in Group 2 the regression in the same period equaled 0.59 ($SE = 0.03$); a non-significant difference ($t = 0.20$, $p = .42$). Hence, we cannot show that the effect of the intervention depended on students' writing proficiency.

Social validity of the intervention

In the logbooks the participating teachers indicated that they highly appreciated the lessons: on a scale from 1 to 10, they rated the lessons with an average score of 7.6. Additionally, teachers estimated the level of difficulty of the lessons for their students with an average score of 2.9 on a 5-point scale. This indicates that the lessons were challenging, but not too difficult. Further, teachers reported on the level of difficulty of executing the lessons. This aspect was rated with an average score of 2.2 on a 5-point scale, indicating that teachers felt that they were sufficiently equipped to teach the lessons to their students. Finally, the main reason that teachers gave for not fully completing the program, according to the remarks in the logbooks, is that it was too challenging to fit in the two extra writing lessons into their regular weekly schedule.

DISCUSSION

In this large-scale intervention study we tested the effectiveness of the newly developed writing program Tekster for grade 4 to 6 in a natural setting with teachers delivering the program in their own classroom. This two-month writing intervention program significantly improved the writing quality of students in grade 4 to 6. Students' individual writing quality did not only increase, but also became more consistent over time. The switching replication design allowed us to replicate the intervention within this study. Results show that the intervention was equally effective in both groups. Moreover, we find that students in Group 1 still wrote qualitatively better texts at the delayed posttest measure than at the pretest measure, indicating that the effect of the intervention was maintained after two months. Although there was a significant improvement of students' writing scores in all grades, the effect of the intervention was slightly smaller in grade 6 than in grade 4 and 5. Further, results show that girls outperformed boys on all measurement occasions, but that the effect of the intervention was the same. Lastly, results of an aptitude treatment analysis showed that the effect of the intervention did not depend on students' writing proficiency.

Effect size of the intervention

The effect size of the intervention on students' writing was moderate (ES = 0.32). This effect size is based on the average of completed lessons (which was ten) and is therefore a conservative estimation of the real effect. Results showed a dosage effect, which means that the more lessons students completed, the more their scores improved. If the effect would have been based on scores of students who completed all lessons, the effect size would increase from 0.32 to 0.40. However, in a general educational setting, it is not unusual that a number of students does not complete all lessons due to absence or other priorities. Further, the scheduling of two lessons a week appeared to be too ambitious for a large number of participating teachers. In regular writing education, only 45 minutes per week are devoted to writing at best (Henkens, 2010; Rietdijk et al., 2015). During the intervention period, teachers had to devote twice as much time to teaching writing. This extra teaching time had to be accomplished on top of the regular curriculum, which led to scheduling conflicts. Given these circumstances, it is promising that, on average, teachers succeeded in doing more than one lesson a week. Taking these circumstances into account, it is more realistic to base the effect size on the average number of completed lessons. On the basis of the dosage effect, we expect that students will make more progress when they complete the whole program. This can be achieved more easily when Tekster is spread out over a longer period of time (e.g., one lesson a week), and/or when the program would contain more lessons. Further research is needed to get more insight in this aspect.

The effect of the intervention can also be interpreted in a more intuitive way by comparing it to the general improvement in writing skills of students between grade 4 to 6 (Lipsey et al., 2012). The results in the present study showed that students in fifth or sixth grade wrote texts of higher quality than fourth grade students; the average improvement in text quality scores was 8.07 points per grade. Hence, after following only two months of the newly-developed writing program Tekster, students' writing improved by more than half a grade.

Generalizability of the results

In comparison to similar strategy-focused intervention studies aimed at grade 4 to 6 in a general educational setting, the effect size of this study (0.32) is notably smaller (cf. Graham et al., 2012; Koster et al., 2015, average ES 1.02 and 0.96 respectively). However, in contrast to most other intervention studies, Tekster was tested on a very large scale: involving 1420 students from 60 classes from 27 schools. Moreover, whereas most intervention studies used only one task as an indication of the effectiveness of their writing program, we tested students' overall writing proficiency with nine writing tasks in three genres: narrative, persuasive, and descriptive. Effects are therefore not only generalizable over students, but also over teachers and tasks. If we would neglect the variance component related to tasks and classes, the effect size of our intervention increases to 0.63, and to 0.80 if the full program would have been completed. This is more in line with effect sizes reported in other intervention studies.

Maintenance effects

Our results show that students' writing quality is still significantly above pretest level two months after the end of the program, which suggests that the intervention induced a lasting change in students' writing and teachers' instructional practice. However, we also see that students did not continue to gain after the end of the intervention period. This seems contrary to expectations, as students have learned an overall strategy for writing that can be applied to any writing assignment, and teachers have learned to optimize their writing instruction. During the intervention period, students and teachers engaged more and more intensively in writing activities than they normally did. It is hardly surprising that when teachers and students returned to the existing writing activities and routines again, the upward trend flattened, as there are far less learning opportunities provided. Henkens (2010) showed that in the average Dutch classroom teachers only provide writing instruction twice a month and that this does not lead to any significant improvement in students' writing. In the present study this is illustrated by the fact that students in the control group (i.e., Group 2 between the first and the second measurement occasion) do not show any improvement in writing quality.

It should be noted, however, that conclusions about the maintenance effect of the intervention are true only under the assumption that tasks were equally difficult and the effect of the intervention (i.e., interaction between condition and time) was the same for students in both conditions. Although we have tried to keep the writing tasks as similar as possible over the three measurement occasions, used the same rating procedure in which raters used the same benchmark scale for equal tasks across occasions, and averaged over three writing tasks per occasion, we cannot entirely exclude the possibility that differences or similarities between scores over time (within conditions) are due to coincidence.

Grade differences

Although Tekster was overall effective in improving students' writing performance, results show that students' writing quality in grade 4 and 5 improved slightly more than the writing quality of sixth grade students. An explanation for this can be that, even though the general approach is the same across grades, the acronyms slightly differ. Grade 6 is the only grade in which students are explicitly instructed to evaluate and revise. Research has shown that revising is difficult for students. In order to be able to revise, students require awareness of the goals and audience of the text, they have to be able to critically read and evaluate their text, and they have to know how they can fix problems, on local level as well as textual level (Fitzgerald, 1987). Ideally, students start working with Tekster in grade 4, focusing on pre-writing strategies and gradually move on to grade 6, with the focus on revision. As this experiment was a cohort study, grade 6 students lack the basics that were the focus of instruction in grade 4 and 5. We have addressed this issue by creating overlap in the topics that are covered in the different grades, but it might be that learning this overall approach at once was more complicated for students than the simpler versions of the acronym that were used in grade 4 and 5. A longitudinal study would provide more insight in this matter.

A longitudinal study of Tekster would also shed more light on the learning trajectory of students across grades. The Dutch Inspectorate of Education (Henkens, 2010) reported that at present students hardly progress in their writing from grade 4 to grade 6. As we have developed a systematic approach for the teaching of writing in the upper primary grades, we would expect a more continuous development of students' writing performance across the grades as a result.

Effectiveness of Tekster for Learning to Write

Results did not show an aptitude treatment interaction, indicating that all students, less proficient as well as proficient writers throughout grade 4 to 6, benefited from the program to the same extent. This suggests that the program addresses the needs of all students, which is promising, given that in a general classroom students differ considerably in their needs and abilities (Harris et al., 2012). The effectiveness of the program for different types of students can be explained in at least three ways. First, Tekster aims at reducing the cognitive overload during writing by providing students with skills and knowledge to regulate their writing process. Second, the program addresses the double challenge of writing and learning to write at the same time. Third, through the multifaceted approach of Tekster, all students, weak as well as proficient writers, are provided with ample learning opportunities, for example by including coping as well as mastery peer modeling (Braaksma, 2002). That Tekster enhances the performance of all students is promising for whole classroom use, as in a typical upper elementary classroom the whole range of abilities will be represented.

It should be noted that, although the program as a whole was effective in improving the writing performance of students, we cannot make claims about the effectiveness of the individual components. We simply do not know which component is the most powerful ingredient of our approach. What we do know from previous research is that the combination of strategy-focused instruction and observational learning is highly effective in improving students' writing performance (Fidalgo et al., 2015). Fidalgo and colleagues assessed the effectiveness of four different instructional components of a strategy-focused writing training: modeling and reflection, direct instruction, peer feedback, and solo practice for sixth grade students, by manipulating the instructional sequence. Their results show that all positive effects are particularly related with the modeling and reflection component. The way our study is designed does not allow for any conclusions regarding the effect of the observational learning component, but based on the findings of Fidalgo and colleagues (2015), it could be that, especially in combination with strategy-focused instruction, modeling may have contributed substantially to the effectiveness of our program. Additional research is needed to examine this to a further extent.

Teachers' implementation of Tekster

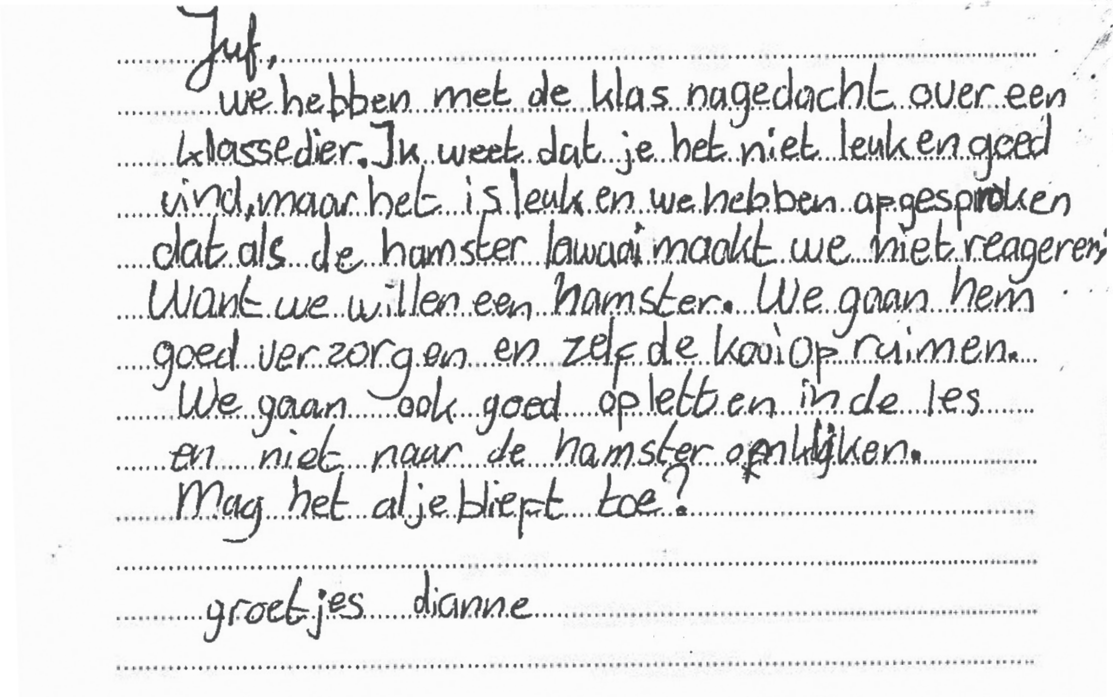
The writing program was delivered at class-level by regular teachers in their own classrooms. Teachers from a large variety of schools participated in this study. Although this contributes considerably to the ecological validity of this study, differences between classes are maximized. Differences between classes are also caused by differences between teachers, such as teaching experience, background, teaching styles and individual preferences (Hattie, 2009). In previous studies, researchers often controlled for the differences between teachers by implementing the intervention themselves (cf. Gordon & Braun, 1986; Kellogg, 1988) or by intensively training teachers or teacher assistants to implement the intervention (cf. Fidalgo et al., 2015; Graham et al., 2005). Whereas intensive training is possible in a relatively small-scale study of one or two classes, this is not a feasible option when an intervention is implemented on a large scale.

Although this study showed large differences between teachers, the results also showed that differences between teachers were reduced after the intervention. This suggests that teachers have adapted their instructional practice as a result of the program. This seems to be confirmed by the fidelity measures, which reveal that teachers closely adhered to the lesson plans as indicated in the manual, and that they applied the key components of the intervention program, i.e., modeling, the acronym, and the steps of the strategy. Further, social validity measures indicated that teachers highly appreciated the program. The lesson plans in the manual facilitated the preparation as well as the execution of the lessons. In addition, they perceived the lesson program as challenging, but not too difficult.

It is promising that already after a limited amount of training, teachers were capable to apply the key components in their instruction. However, whereas classroom observations provide information on what was done during the lessons, the observational data do not allow for statements on the quality of the executed lessons. In further research it is necessary to observe not only what teachers do in class, but also how they do this, for instance by videotaping and subsequently analyzing lessons to get a clear picture of the practice of every teacher and whether and how teachers adapted the program to their own practice.

General conclusion

To conclude, this study has shown that an overall approach in which several evidence-based instructional practices for teaching writing are combined is effective in improving elementary students' writing quality. This study is unique for the following reasons. First, through a switching replication design we were able to replicate the effect within one study, with the same results. Together with the scale of the study with a large sample of Dutch schools, this allows us to make robust claims about the effectiveness of the writing program. Second, in this study we examined the impact of the writing program in a naturalistic setting, as the intervention was delivered in 60 intact classes by regular teachers, after having received only a short training. Third, students learn a general strategy for writing, irrespective of genre, and the quality of their writing was measured with multiple writing tasks using multiple text types. It is therefore possible to generalize the results to overall writing proficiency in a general educational setting. All in all, this study demonstrates that a comprehensive writing program, such as Tekster, is a promising approach to improve elementary students' writing.



Juf,
we hebben met de klas nagedacht over een
klassedier. Ik weet dat je het niet leuk en goed
vind, maar het is leuk en we hebben afgesproken
dat als de hamster lawaai maakt we niet reageren;
Want we willen een hamster. We gaan hem
goed verzorgen en zelf de kooi op ruimen.
We gaan ook goed opletten in de les
en niet naar de hamster omkijken.
Mag het aljeblieft toe?

groetjes dianne

Chapter 6

A LETTER OF ADVICE: THE RELATIONSHIP BETWEEN METACOGNITIVE KNOWLEDGE AND WRITING PERFORMANCE FOR ELEMENTARY STUDENTS

This chapter is based on: Koster, M., Bouwer, R., & Van den Bergh, H. (2016). A letter of advice: The relationship between metacognitive knowledge and writing performance for elementary students. Manuscript is submitted for publication.

This study examined the relationship between metacognitive knowledge about writing and writing performance, and whether the writing performance of students in upper elementary grades could be improved through a writing intervention aimed to enhance students' declarative, procedural, and conditional knowledge of writing. Writing knowledge was examined by means of a letter of advice to a fictitious peer. A quasi-experimental posttest-only design was used with two conditions: a control condition (*N*=373) in which students wrote the letter without any extra instruction; and an experimental condition (*N*=220) in which students first received the intervention before writing the letter. Results show that students' knowledge primarily pertains to lower order aspects of writing: punctuation and capitals, and spelling and grammar. The intervention has increased students' knowledge of writing, as students in the experimental condition gave more writing advice than students in the control condition. Further, students in the experimental condition gave more process and organizational advice than students in the control condition. Concerning text quality, we found that students in the experimental condition consistently outperformed students in the control condition. Regarding the relationship between writing knowledge and writing performance, we found that five categories of writing knowledge contributed positively to text quality: punctuation and capitals, spelling and grammar, presentation, organization, and process. This study demonstrates that there is a relationship between metacognitive knowledge about writing and writing performance, and that writing performance can successfully be improved by enhancing students' declarative, procedural, and conditional knowledge about writing.

INTRODUCTION

Together with reading and arithmetic, writing is one of the most important basic skills students need to master during the period they spend in elementary school. Adequate writing skills are increasingly important, due to the rapid digitalization that has taken place during the last two decades. Communication that used to be face-to-face or oral is more and more being replaced by e-mail and text messaging, requiring well developed reading and writing skills. Further, it has been established that writing performance is an important predictor of both academic and professional success (National Commission on Writing, 2003). It is therefore essential that students learn the basics of written communication already at a very young age.

Despite the fact that writing skills are of vital importance, consecutive large Dutch national assessment studies (Krom, Van de Gein, Van der Hoeven, Van der Schoot, Verhelst, Veldhuijzen & Hemker, 2004; Kuhlemeier, Van Til, Hemker, De Klijn, & Feenstra, 2013) showed that a majority of students at the end of elementary school (grade 6) are not able to compose texts that sufficiently convey a simple message to a reader. The Dutch Inspectorate for Education found that in the average classroom attention and time devoted to writing are limited, and that the majority of teachers do not succeed in effectively teaching writing (Henkens, 2010). An explanation for these shortcomings in writing education is that writing is a neglected topic in teacher education: during their own professional development teachers hardly received any training in writing or in the teaching of writing, whereas teaching writing requires a high degree of a teacher's own writing skills, knowledge of and insight in the writing process (Van der Leeuw, 2007). Consequently, in their daily classroom practice, teachers focus on the aspects of writing that they do have knowledge about, such as the written product and aspects thereof (writing topic and text type, but also spelling, grammar and punctuation), instead of providing their students instruction and support concerning aspects of the writing process, promoting prewriting or revising activities, and student collaboration during writing (Franssen & Aarnoutse, 2003; Henkens, 2010; Pullens, 2012). It can be concluded that teachers are not sufficiently equipped to assist their students during the writing process. In his report Henkens (2010) presented several suggestions to improve writing education. His most salient suggestion was to shift the focus from the writing product to the writing process in the teaching of writing, and to facilitate this process approach to writing by providing teachers with the requisite knowledge, skills and material.

The starting point for the shift from a product-directed approach to a more process-directed approach to writing was the cognitive process model of writing by Flower and Hayes (1980, 1981). Based on thinking aloud protocols of experienced writers, they constructed a model of the observed thinking processes and constraints. In their model, Flower and Hayes distinguished three basic writing processes: planning, translating and reviewing, which are all controlled by a monitor. It is emphasized in this model that writing is a recursive process; during writing these processes do not typically occur in a fixed sequence, but are interwoven with each other. However, as the model of Flower and Hayes is based on the writing process of experienced (expert) writers, it is less suitable to describe the writing process of young and inexperienced writers. Bereiter and Scardamalia (1987) examined the writing processes of beginning writers. Contrary to more experienced writers, beginning writers are still engaged in acquiring the skills that are fundamental for writing, such as handwriting, spelling, grammar, and punctuation, which leaves less cognitive capacity for the higher-order aspects of writing. To manage the cognitive overload of having to deal with several cognitive activities simultaneously, young and inexperienced writers predominantly adopt a "knowledge-telling" strategy. Simply stated, they generate ideas, write them down, generate more ideas, write those down as well, and so on. The knowledge-telling strategy is characterized by a lack of planning and structuring, which seldom leads to a good text (Bereiter & Scardamalia, 1987).

A more mature writing strategy is "knowledge-transforming", characterized by reworking of thoughts and interaction between text processing and knowledge processing (Bereiter & Scardamalia, 1987). During knowledge-transforming, ideas are selected, structured and presented with the intended communicative goal and audience in mind, which leads to texts of a better quality than generally is the case when using the knowledge-telling approach. Students who predominantly employ a knowledge-telling strategy can learn to master knowledge-transforming procedures through effective instruction, after the lower-level skills, that are conditional to writing, have become more automatized (Bereiter, Burtis & Scardamalia, 1988). When the lower-level skills are automatized, students have more memory capacity available for higher order processes like planning and revising. Thus, to improve the writing proficiency of beginning writers, writing instruction has to aim to bridge the gap between knowledge-telling and knowledge-transforming.

In process models of writing the importance of content knowledge as well as process knowledge in the writing process is recognized by the inclusion of a knowledge component in the models (Bereiter & Scardamalia, 1987; Flower & Hayes, 1981). In their model, Flower and Hayes (1981) locate the knowledge component in long-term memory, and similar to other knowledge stored in long-term memory, this knowledge is retrieved automatically by cues in the task situation, or it can be activated by deliberate searching (Flavell, 1979). The nature of this knowledge and its role in the writing process is still indistinct. However, it might very well be that this knowledge component is a crucial factor in learning to write. McCutchen (1986), for instance, has demonstrated that more topic knowledge leads to more coherent texts, but she also observed that in text production there is also more generalizable knowledge about writing and texts involved. Thus, to be able to write well, a writer must not only know what to write, but he must also know how to write it (McCutchen, 1986; Schoonen & De Glopper, 1996). This last kind of knowledge, "knowing how", is referred to as metacognitive knowledge (Flavell, 1979; Harris, Graham, Brindle, & Sandmel, 2009). Paris, Lipson and Wixson (1983) distinguished three types of metacognitive knowledge: declarative knowledge (information about task characteristics, such as the writing topic, genre, structure and goal of the task), procedural knowledge (information about the execution of various actions and strategies to attain specific goals), and conditional knowledge (knowing when and how to apply strategies in order to attain the intended goal). So, a proficient writer knows what he wants to write and can select and apply the appropriate strategy to write a text that meets the intended goal.

Advocates of a process-directed approach to writing instruction assume that there is a direct relationship between the (type of) metacognitive knowledge a writer possesses and text quality. Writers who have more metacognitive knowledge know on which aspects they need to focus during writing and devote more attention to planning and structuring, which leads to higher text quality (Deane, Odendahl, Quinlan, Fowles, Welsh, & Bivens-Tatum, 2008). Several studies investigated the relationship between students' knowledge of writing and text quality, and their findings demonstrate that students with more metacognitive knowledge indeed write better texts (Englert, Raphael, Fear, & Andersen, 1988; McCutchen, 1986; Saddler & Graham, 2007; Schoonen & De Glopper, 1996; Schoonen, Van Gelderen, Stoel, Hulstijn, & De Glopper, 2011; Trapman, Van Steensel, Van Schooten, Van Gelderen, & Hulstijn, 2012). Further, these studies show, without exception, that more proficient students focus more on the higher-order aspects of writing, such as structure and style, instead of lower-order aspects, such as handwriting, spelling and grammar. Thus, it is warranted to conclude that there is indeed a relationship between metacognitive knowledge and text quality. However, we must bear in mind, that this is correlational evidence: to examine the causality of this relationship further (experimentally controlled) research is required.

Research into students' metacognitive knowledge about writing demonstrates that in the elementary grades students already possess metacognitive knowledge about writing, but that this knowledge is predominantly declarative. It has been shown that children at a very young age (grade 1 and 2) are capable to understand the communicative nature and purpose of writing (Kos & Maslowski, 2001; Olinghouse & Graham, 2009; Shook, Marrion, & Ollila, 1989), but research also demonstrates that these young students reckon they need more practice at mechanical aspects of writing to become better writers (Kos & Maslowski, 2001; Shook, Marrion, & Ollila, 1989). This focus on the appearance of writing (handwriting, spelling, punctuation, and neatness) persists in higher elementary grades (Barbeiro, 2011; Lin, Monroe, & Troia, 2007; Wray, 1993), but from grade 5 onwards students show procedural and conditional knowledge of writing: they are able to vary cognitive strategies depending on context, purpose, and genre of the writing assignment, and display understanding of audience awareness (Barbeiro, 2011; Gillespie, Olinghouse, & Graham, 2013; Lin et al., 2007; Olinghouse, Graham, &, Gillespie 2014). Although in secondary school students still consider neatness and spelling important aspects of writing, they are also aware of the higher-order aspects of writing, such as structure, content, and style (Braet, Moret, Schoonen, & Sjoer, 1993; Crismore, 1982; Schoonen & De Glopper, 1996).

Various intervention studies have tried to expand students' knowledge on (aspects of) writing with the underlying assumption that increased knowledge would result in an improvement in students' writing performance. Raphael, Englert, and Kirshner (1989), and Englert, Raphael, Anthony, and Stevens (1991) investigated the effect of writing intervention programs on metacognitive knowledge in grade 4-6 and found significant improvement in metacognitive knowledge in the areas focused on, such as communicative context and text structure, as well as a transfer effect to other content areas. However, they did not include a measure of text quality in their study, therefore it is unknown whether increased metacognitive knowledge resulted in an improvement of students' writing performance.

A large number of intervention studies have focused on improving students' writing performance by increasing students' declarative, procedural and conditional knowledge through the teaching of strategies for writing. Several meta-analyses (Graham, 2006; Graham, McKeown, Kiuhara, & Harris, 2012; Graham & Perin, 2007; Koster, Tribushinina, De Jong, & Van den Bergh, 2015) have shown that explicit strategy instruction is a very effective type of intervention to improve the quality of students' writing, and that it is even more effective when combined with teaching students self-regulation skills. The most dominant (and by far the most investigated) example of this combination is the SRSD (Self-Regulated Strategy Development) approach to writing (cf. Graham, Harris, & Mason, 2005) or variations thereof (cf. Bouwer, Koster, & Van den Bergh, 2016a; Torrance, Fidalgo, & Robledo, 2015). In self-regulated strategy instruction writing strategies, such as planning, organizing, revising, and self –regulation strategies, such as goal-setting and monitoring the progress toward these goals, are explicitly taught, as well as the declarative, procedural and conditional knowledge that is needed to successfully carry out a writing task. These strategy-focused intervention studies have demonstrated to improve students' text quality, but did not measure the change in metacognitive knowledge. It is therefore unknown whether in these studies an increase in metacognitive knowledge has led to improvements in text quality.

So far, only a couple of studies have tried to bridge the gap between correlational research on metacognitive knowledge and text quality and writing intervention research, which allows for claims about a possible causal relationship between metacognitive knowledge and writing performance (Graham, Harris, & Mason, 2005; Harris, Graham, & Mason, 2006). These studies were conducted with students in the early elementary grades: grade 3 and 2, respectively. Both these studies show that a strategy-focused intervention focusing on planning and writing stories has a positive impact on the writing performance and knowledge of students. Although students in grade 2 were more knowledgeable about writing, especially about the writing process, and wrote longer and more complete stories, there was no significant difference in text quality in comparison to a control group (Harris et al., 2006). It could be that these students were too young, or may have needed more practice. On the other hand, Graham and colleagues (2005) showed that strategy-focused instruction does not only have an impact on students' knowledge of writing (after the intervention students focus more on substantive processes instead of on production aspects), but also on students' writing quality, compared to a control condition. To gain more insight into the relationship between metacognitive knowledge and writing performance, the development of these aspects across grades, and whether (and how) students' writing performance can be improved by enhancing their metacognitive knowledge about writing knowledge, more research is needed, especially in the upper elementary and secondary grades.

The present study examines the relationship between knowledge about writing and writing quality in students in the upper elementary grades, i.e., grade 4-6. As well as examining the relationship between knowledge about writing and writing quality in elementary students, we also investigate whether the writing performance of students can be improved by increasing their knowledge of the writing process. This combination of correlational research on metacognitive knowledge and text quality and writing intervention research allows for claims about a possible causal relationship between metacognitive knowledge and writing performance. The intervention addresses students' declarative, procedural, and conditional knowledge of writing. Students are taught an overall strategy to approach writing tasks, based on the steps of the writing process (generate content, organize content, reread, evaluate, revise), as well as the procedural and conditional knowledge that they need to successfully apply this strategy to writing tasks. We investigate the effect of the intervention on (a) knowledge of writing of students grade 4 to 6, (b) text quality, and (c) the relation between writing knowledge and text quality.

METHOD

Participants

For this study 26 classes grade 4 to 6, from 25 schools were recruited from various parts of the Netherlands. In total, 593 students (48.6% male, 51.4% female) participated in the study. Participants' age ranged from 8 to 13 years, with a mean age of 10.6 years ($SD = 1.05$). Classes were assigned to one of two conditions: control ($N = 18$ classes) or experimental condition ($N = 8$ classes). Students in the control condition were given a writing assignment without any specific instruction; students in the experimental condition first received a series of three writing lessons before accomplishing the same writing assignment. Table 6.1 gives an overview of the number of participants per condition per grade.

Procedure

To measure students' knowledge of writing, writing conventions, and the writing

process, students had to write a letter of advice to a fictitious peer on how to get good grades for writing. More specifically, the assignment provided an overview of the aspects of writing that students consider important. Additionally, the letters were the measure of students' writing performance.
The assignment was phrased as follows:

> "Next week, a new student will arrive in your classroom: Like. Like was born in the Netherlands, but has lived in England for some time. The school system in England is different from the Dutch system. Like is not exactly sure how to write a good text in Dutch, because there are many aspects that you have to take into account. Write a letter to Like to explain how to write a good text in Dutch, and thus, to get good grades for writing. Give Like as many tips and advice you can think of."

We purposely choose a gender neutral, unusual name for the new student (Like), to avoid biased students' reactions. Further, the assignment was deliberately vague on the kind of expected advice, in order to evoke as many types of advice as possible.

In the control condition, the assignments were administered by (under-) graduate students. In the experimental condition, the assignments were administered by the classroom teacher. In both conditions, no extra instruction was given regarding the assignment. There was no time limit for students: they could take as much time as they needed to finish the task. Students were not allowed to interact during writing.

Table 6.1 Overview of number of participants per grade

Grade	*N* Control	*N* Experimental	Total
4	130	71	201
5	117	61	178
6	126	88	214
Total	373	220	593

Writing knowledge

To measure what knowledge students possess about writing, the writing advice that was given by the students was coded, using a simplified and adapted version of the coding scheme of Schoonen and De Glopper (1995). Writing knowledge was classified into two main categories, declarative knowledge (information about task characteristics) and procedural knowledge (information about the execution of action and strategies). Measuring conditional knowledge (information on when and how to apply strategies) lies beyond the scope of the present study, as measurement of this type of knowledge requires process measures, such as think aloud protocols. The category declarative knowledge contains advice concerning punctuation and capitals, spelling and grammar, presentation, organization, content, and style. The category procedural knowledge contains advice concerning process aspects of writing. Advice that did not fit into these categories was classified as 'other', this was advice concerning behavioral or physical aspects of writing, or advice unrelated to writing (miscellaneous). Table 6.2 provides an overview of the categories, with one example of advice per category.

The classification of students' advice in the different knowledge categories was done by two raters. A random selection of 120 essays (approximately 20% of the total amount) was coded by both coders. The reliability between raters over all categories was high, Cohen's κ = .87.

Table 6.2 Overview of advice categories, with examples

Category	Example
Declarative knowledge	
Punctuation & Capitals	Always start your sentence with a capital, end with a full stop.
Spelling & Grammar	Remember to apply the spelling rules.
Presentation	Write neatly.
Organization	Start your story with an introduction.
Content	You need a good topic to write about.
Style	A story must not be boring, make some jokes.
Procedural knowledge	
Process	You have to make a draft first.
Other	
Behavior	You should pay attention to what the teacher says.
Physical	Sit straight and hold your pen the right way.
Miscellaneous	Learn the alphabet.

Text quality

To assess text quality, a continuous scale with five benchmark essays was used. The center position on this scale was an average essay that was assigned an arbitrary score of 100 points, the other essays on the scale were one (115 points) and two (130 points) standard deviations above average, and one (85 points) and two (70 points) standard deviations below average. The rating scale can be found in Appendix J. Raters compared the students' essays to the benchmark essays and administered each essay a score accordingly. Raters were trained in advance on how to apply the scale. Each essay was rated three times by overlapping rating teams (Van den Bergh & Eiting, 1989). The text quality score of each text was obtained by averaging the scores of three raters. There were 12 raters in total, and each rater rated overlapping portions of essays (around 150 in total). The reliability for each rater was assessed as well as the reliability of each jury. The average reliability of the juries was high, $\rho = .89$, varying from $\rho = .85$ to $\rho = .93$.

Writing intervention

For each grade we developed a series of three writing lessons addressing the three types of metacognitive knowledge distinguished by Paris, Lipson and

Wixson (1983). The lessons aimed to increase students' declarative knowledge by explicitly teaching text structure, communicative goals for writing, and genre specific aspects. Students' procedural and conditional knowledge was enlarged by teaching them strategies to approach writing tasks as well as teaching them how and when to use this strategy.

The strategies that were taught were based on the steps of the writing process with acronyms serving as a mnemonic: animal names of which each letter represents a step of the process (see also Bouwer, Koster, & Van den Bergh, 2016a; Koster, Bouwer, & Van den Bergh, 2014a, 2014b, 2014c). These animals (VOS (fox) for grade 4, DODO (dodo) for grade 5, and EKSTER (magpie) for grade 6[1] served as the overarching theme in the lessons. The main focus in grade 4 was on prewriting activities, this shifted to postwriting activities in grade 6. Students' conditional knowledge was further increased by practicing in applying the strategy to various authentic writing tasks with clear communicative goals. In grade 4, the students had to apply the strategy to writing a postcard, an interview, and an instructive text on how to fold a paper airplane. In grade 5, students had to write a letter of complaint, an interview, and instructions how to play a game. In grade 6, students wrote a letter of complaint, a recipe, and an instruction for a craft project.

The 60-minute-lessons were scheduled once a week during a three-week period. The lessons were taught by the regular classroom teacher. For each lesson a detailed lesson plan was included providing an overview of the instruction and activities of the lesson with a time planning for each phase of the lesson. These lesson plans described the activities the teacher was expected to execute during the lesson and provided suggestions when to use modeling during instruction. Students worked on printed worksheets: ample writing space was provided for each phase of the lessons: for prewriting, writing, and postwriting activities. At the end of the intervention the students wrote the letter to Like (see above).

Intervention fidelity

Worksheets. The worksheets of all lessons of students were collected after the intervention to check if they completed all phases of all lessons, including pre- and postwriting activities. Analysis of the worksheets showed that all students completed all phases of all lessons.

Observations. To check how teachers implemented the intervention, observations were conducted in 50% of the classrooms. In these randomly selected classrooms all three lessons were observed. Our observation instrument was designed to map teacher's activities during the lessons. Every 30 seconds was

[1] VOS (fox) stands for Verzinnen (generate content), Ordenen (organize), Schrijven (write); DODO (dodo) stands for Denken (think), Ordenen (organize), Doen (do), Overlezen (read); Ekster (magpie) stands for Eerst nadenken (think first), Kiezen & ordenen (choose & organize), Schrijven (write), Teruglezen (reread), Evalueren (evaluate), Reviseren (revise).

tallied whether the teacher was on task or off task: on task if the teacher was executing the actions as specified in the lesson plan for that phase of the lesson, off task if the teacher was involved in other activities than teaching writing, such as fetching a cup of coffee or talking to a colleague. Further, observers had to register whether the teacher referred to the acronym and/or explained (the steps of) the strategy, which were the key elements of the intervention. Each classroom was observed by one undergraduate student, there were five observers in total. To optimize observers' agreement, all observers were trained in advance.

Analysis of the observational data showed that teachers were on task on average 75% of the total lesson time. Teachers' off-task behavior was mostly observed at the end of the lessons (during writing and postwriting), when students were independently working on their texts. Observed off-task behavior involved reading other material, checking e-mail, or preparing the next activity or lesson. Lastly, in every observed lesson but one, teachers referred to and explained the acronym and the strategy. The lesson in which the strategy and acronym were not mentioned was the third lesson in a grade 5 lesson: it might be that the teacher reckoned that students were familiar with the strategy by then.

Data analysis

As students were nested within classes, the data were analyzed by applying multilevel modeling. To analyze students' writing knowledge we fitted four models to the data, by systematically adding parameters. The first model (Model 1a) was our basic model, in which we included the distinction between categories. Next, we examined whether there were differences in the amount of students' writing knowledge between categories, depending on students' grade (Model 1b). Further, we investigated whether students' writing knowledge differed between categories, depending on condition (Model 1c). Lastly, we tested whether the amount of knowledge varied among categories as a result of the combination of grade and condition (Model 1d)[2].

To analyze students' text quality scores four models were fitted. First, a basic intercept model was fitted to the data (Model 2a), in which we accounted for the variance in text quality between classes and between students. Subsequently, by systematically adding parameters, we examined whether text quality scores differed between grades (Model 2b) and between conditions (Model 2c). In the final model we investigated the interaction effect of grade and condition (Model 2d).

To examine the relation between the different categories of writing knowledge and text quality, we fitted four models to the data. The starting point for model fitting (Model 3a) was the best fitting model for text quality, in the second model (Model 3b) we added the categories of writing knowledge to the model, to test whether the relation between text quality and writing knowledge was different between categories. Next, by systematically adding parameters, it was tested

[2] We did not estimate main effects for grade and condition for the categories as this assumes that differences between categories are the same for different grades or conditions, which is highly implausible.

whether this relation differed among grades (Model 3c), and whether the relation between text quality and writing knowledge was different between conditions (Model 3d). In the last model was examined whether there was an interaction effect between grade and condition for the relationship between text quality and writing knowledge (Model 3e).

RESULTS

Writing knowledge

In total 593 letters were included in the analysis. The total amount of advice was 3656, an average of 6.2 ($SD = 3.36$) per letter. In both conditions, the variation between letters was huge: in some letters barely any advice was given, whereas others contained over ten tips. Most advice was related to the written product and merely dealt with surface features. Punctuation and Capitals was the largest advice category, in the control condition as well as the experimental condition (resp. 27% and 29% of the total number of advice per condition). In the control condition the second largest category was Spelling and Grammar (21%), followed by advice concerning Behavior (16%). In the experimental condition, however, the second largest category was advice concerning Process (15%), followed by Spelling and Grammar (14%). In the other categories, these percentages ranged from 1% (Style) to 11% (Presentation). The letters varied considerably in the types of advice that was given: in some letters students focused exclusively on one type of advice, while in other letters students gave various types of advice.

Table 6.3 shows the results of the fit and comparison of the four multilevel models that were fitted to the data. First, we estimated the differences in the amount of writing knowledge between categories (Model 1a). Subsequently we examined whether there were differences between grades in the amount of writing knowledge in the different categories by adding the interaction effect of category and grade to the model. This led to a significant improvement of the model (Model 1a versus Model 1b, χ^2 (20) $= 46.61$, $p < .001$), indicating that there were indeed differences in the amount of writing knowledge depending on grade.

Next was tested whether there were differences between conditions in the

Table 6.3 Comparison of multilevel models for writing knowledge

		-2 Log	Comparison			
Model	N_{pars}	Likelihood	Models	ΔX^2	Δdf	p
1a differences between categories	30	15758.18				
1b + grade*category	50	15711.57	1a vs 1b	46.61	20	< .001
1c + condition*category	60	15669.90	1b vs 1c	41.67	10	< .001
1d + grade*condition *category	80	15624.18	1c vs 1d	45.72	20	< .001

amount of advice that was given in the different categories by testing the interaction effect of category and condition. This interaction effect was also significant, χ^2 (10) = 41.67, $p < .001$ (Model 1b versus 1c), thus, there were differences between conditions in the amount of advice that was given in the various categories depending on condition.

In the last model (Model 1d) we tested condition, by testing the interaction effect of grade, condition, and category. This model provided the best fit for the data, χ^2 (20) = 45.72, $p < .001$ (Model 1c versus 1d), indicating that there were differences between categories in the amount of given advice and that these differences depended on the combination of grade and condition.

To facilitate the interpretation of the interaction effects, they are explained using a graphical overview (Figure 6.1), that provides the average amount of knowledge for each category per grade and per condition, as well as the 95% confidence interval. As can be seen, the means varied considerably between categories as well as between grades and conditions. In four categories we found no significant differences between grades or conditions in the amount of advice that was given, these were Punctuation and Capitals, Presentation, Behavior, and Physical. In the other categories we found significant differences between conditions and/or grades. In two categories, Spelling and Grammar and Miscellaneous, students in the experimental condition gave less advice than students in the control condition, but not in all grades. In the category Spelling and Grammar students in grade 6 in the experimental condition gave less advice than their counterparts in the control condition ($p = .007$). In the category Miscellaneous, students in grade 4 in the experimental condition gave less advice than their counterparts in the control condition ($p = .003$).

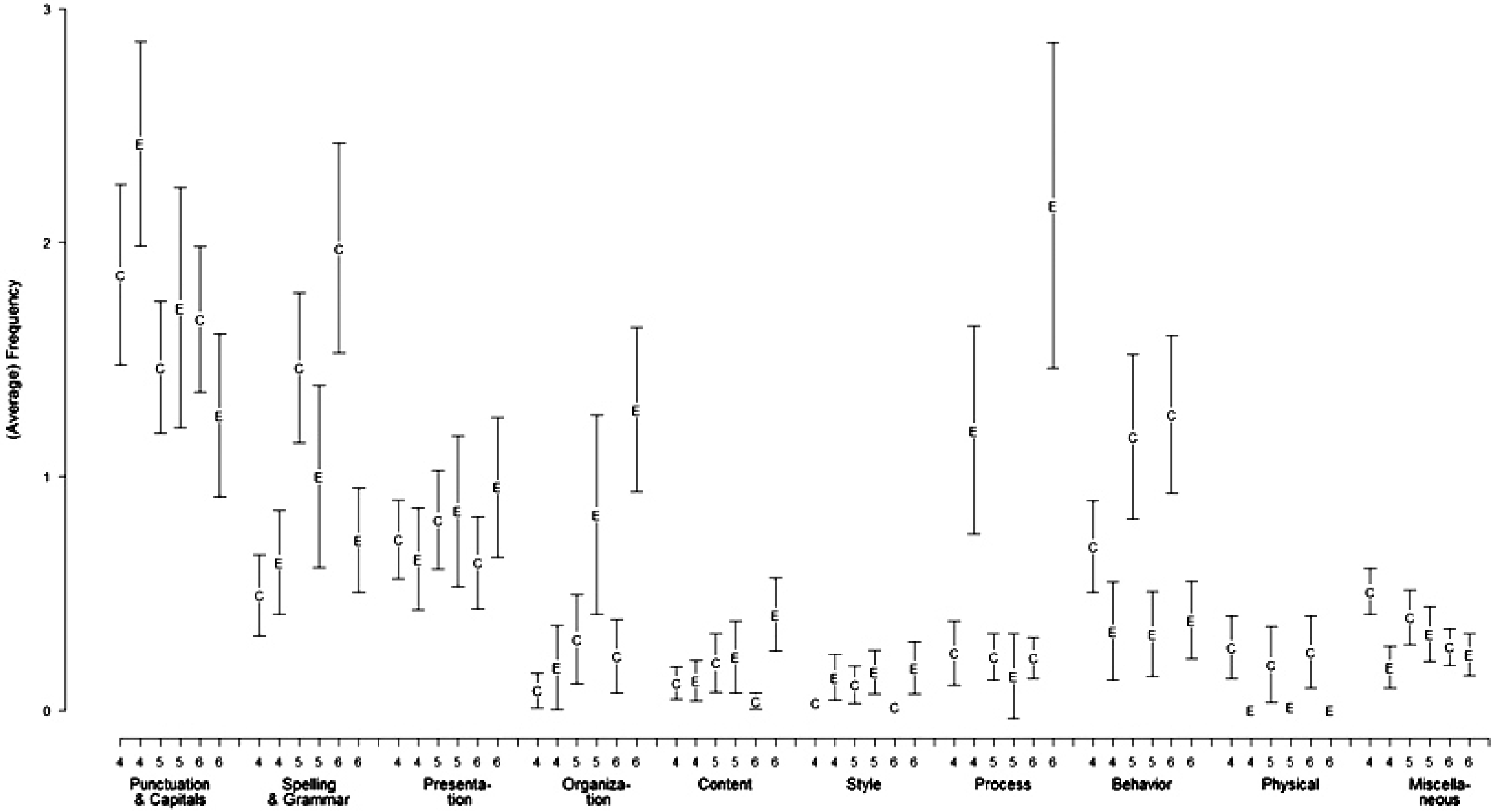
In the remaining categories, in particular in the categories Organization and Process, students in the experimental condition gave significantly more advice than students in the control condition, especially students in grade 6 (resp. $p <$ .001, and $p = .003$). A smaller, but comparable effect was found in the categories Content and Style: in these categories also students in grade 6 in the experimental condition gave significantly more advice than students in grade in the control condition (resp. $p < .001$, and $p = .01$). Additionally we found that in the category Organization also students in grade 5 in the experimental condition gave more advice than their counterparts in the control condition ($p = .04$).

To summarize, we see that students in the control condition provided more advice on mechanical aspects of writing, such as spelling and grammar, and on aspects that are not directly related to writing performance. For students in the experimental condition we see that, especially for students in grade 6, the writing lessons have substantially contributed to their knowledge of the writing process, and to knowledge related to higher order aspects of writing, such as style, content and the organization of texts.

Text quality

To investigate whether students' text quality scores differed between grades and/

Figure 6.1 Overview of the means per category, per grade, per condition, and the 95% confidence interval (C: Control, E: Experimental).



or conditions, and whether these scores varied between grades depending on condition, four multilevel models were fitted to the data. Table 6.4 shows the results of the fit and comparison of the multilevel models. First, a basic intercept model was fitted to the data. Next, we tested whether there were differences in text quality between grades. Results show a significant effect of grade (Model 2a vs. Model 2b, χ^2 (2) = 7.56; p = .02), indicating that there were differences between grades in students' text quality scores. Subsequently we tested whether there were differences in text quality between conditions. Our results show a significant effect of condition (Model 2b vs. Model 2c, χ^2 (1) = 17.17; p < .001), indicating that text quality scores differed between conditions. There was no interaction effect of grade and condition (Model 2c vs. Model 2d, χ^2 (2) = 0.26; p = .88), indicating that the differences between conditions were equal across grades.

Table 6.4 Comparison of models for text quality

			Comparison			
Model	N_{pars}	-2 Log Likelihood	Models	ΔX^2	Δdf	p
2a basic intercept	3	1590.95				
2b + grade	5	1583.39	2a vs 2b	7.56	2	.02
2c + condition	6	1566.22	2b vs 2c	17.17	1	<.001
2d + grade*condition	8	1565.96	2c vs 2d	0.26	2	.88

Table 6.5 displays the parameter estimates and standard error of students' scores for text quality per grade and per condition. Students in grade 6 wrote better texts than students in grade 4 (p = .002), there were no significant differences in text quality between students in grade 4 and 5 (p = .06), and students in grade 5 and 6 (p = .26). Consistently was found that students in the experimental condition outperformed students in the control condition: on average, students in the experimental condition scored 9.99 points higher (SE = 1.99) on text quality than their counterparts in the control condition. The magnitude of this effect was estimated by comparing the effect of the intervention to the total amount of variance (Cohen's d). This resulted in an estimated effect size of 0.72.

Table 6.5 Parameter estimates for text quality scores per condition per grade, standard error between parentheses ($S^2_{student}$ = 177.76, SE = 10.61; S^2_{class} = 15.47, SE = 6.95)

	Control (N=373)	Experimental (N=220)
Grade 4	89.19 (1.65)	99.18 (2.00)
Grade 5	93.14 (1.69)	103.13 (2.01)
Grade 6	95.41 (1.62)	105.40 (1.89)

Relation between text quality and writing knowledge

To examine the relation between writing advice and text quality, five multilevel models were fitted to the data. To facilitate comparison and interpretation, standardized scores were used both for text quality and for the amount of writing knowledge. The results of the fit and comparison of models are displayed in Table 6.6. As the model including the effects of grade and condition on text quality provided the best fit to the data (Model 2c in Table 6.4), we used this model as the starting point for model fitting (Model 3a). Subsequently, we added the writing knowledge categories as parameters (Model 3b). This led to a significant improvement of the model (Model 3a vs Model 3b, χ^2 (10) = 109.66; $p < .001$), indicating that there is a relationship between text quality and writing knowledge. Next, we tested the effect of grade on the relation between text quality and writing knowledge.

Table 6.6 Comparison of models for the relation between text quality and writing knowledge

		-2 Log	Comparison			
Model	N_{pars}	**Likelihood**	**Models**	ΔX^2	Δdf	p
3a grade + condition	6	1566.22				
3b + categories	16	1456.56	3a vs 3b	109.66	10	<.001
3c + categories*grade	36	1412.56	3b vs 3c	44.00	20	<.01
3d + categories*condition	46	1396.20	3c vs 3d	16.36	10	.09
3e + categories*grade *condition	66	1370.80	3d vs 3e	25.40	20	.19

As can be seen, the effect of grade was significant (Model 3b vs. Model 3c, χ^2 (20) = 44; $p = .001$), meaning that the relation between writing knowledge and text quality differed between grades for at least one of the categories. In the next model (3d) we tested the effect of condition, which was not significant (Model 3c vs. Model 3d, χ^2 (10) = 16.36; $p = .09$), which means that we were not able to establish differences in the relation between writing knowledge and text quality between conditions. In the last model we tested whether there were differences in the relation between text quality and writing knowledge depending on the combination of grade and condition, by adding the interaction effect between grade and condition to the model. This effect was not significant (Model 3d vs. Model 3e, χ^2(20) = 25.40; $p = .19$). Thus, there were differences between grades concerning the relation between text quality and writing knowledge, but the relation did not differ between conditions.

Table 6.7 presents the regression coefficients of the different categories of writing knowledge. This table illustrates the relationship between writing knowledge and text quality: five out of ten knowledge categories positively contributed to text quality: Punctuation and Capitals ($p < .001$), Spelling and Grammar ($p <$ .001), Presentation ($p < .001$), Organization ($p < .001$), and Process ($p = .001$). This means that the more students knew about writing (the more advice they give) in these categories, the better the quality of their texts was. Knowledge about Punctuation and Capitals contributed most to text quality: when a student had relatively more knowledge of these aspects (+1*SD*), his text quality score increased by 0.26 a standard deviation. For knowledge in the category Presentation the increase was 0.19 a standard deviation, and for Spelling and Grammar, Organization and Process the increase was 0.16 a standard deviation. It should be noted that the regression coefficients presented in Table 6.7 are the coefficients for grade 6. In three categories we found significant differences between grades concerning the regression coefficients: Spelling and Grammar (grade 4: $\Delta\beta = .24$, $p = .03$; grade 5: $\Delta\beta = .40$, $p < .001$) Presentation (grade 4: $\Delta\beta = .26$, $p = .05$; grade 5: $\Delta\beta = .19$, $p = .02$) and Process (Grade 4: $\Delta\beta = .24$, $p = .007$). As can be seen, all coefficients are positive, indicating that for these grades the effect was even stronger in these categories.

Table 6.7 Standardized regression coefficients and standard errors for categories of writing knowledge on text quality for students in grade 6

Category	Standardized β	*SE*	*t*	*p*
Declarative				
Punctuation & Capitals	0.26	0.04	7.27	< .001
Spelling & Grammar	0.16	0.04	3.80	< .001
Presentation	0.19	0.04	5.41	< .001
Organization	0.16	0.04	4.49	< .001
Content	0.05	0.04	1.53	.13
Style	0.02	0.03	0.54	.59
Procedural				
Process	0.16	0.05	3.48	< .001
Other				
Behavior	0.04	0.04	1.02	.31
Physical	0.02	0.04	0.63	.53
Miscellaneous	0.06	0.04	1.86	.06

Our results show that for students in the upper elementary grades knowledge about lower order aspects of writing, such as punctuation, capitals, spelling, grammar and presentation is important for text quality, but that procedural knowledge also contributes substantially to text quality. Remarkably, advice concerning Content and Style, considered higher order aspects of writing, did not contribute to text quality (resp. $p = .13$, and $p = .59$). It should be noted, however, that these were very small categories; respectively 2.5% and 1.2% of the total amount given advice pertained to these categories.

Finally, three categories of which we logically assumed to have no influence on text quality, i.e., Behavior, Physical, and Miscellaneous did also not contribute to students' writing scores ($p > .05$). On the basis of the advice that was given, we can explain 40% of the variance in text quality, which can be considered a large effect.

DISCUSSION

To bridge the gap between correlational research on metacognitive knowledge and text quality and writing intervention research we examined in the present study, through the combination of a correlational study and an intervention study, the effect of an intervention in which students were taught a process-based writing strategy on writing knowledge, text quality, and the relation between writing knowledge and text quality. The intervention addressed students' declarative, procedural, and conditional knowledge of writing.

Our results show that the intervention has led to an increase in writing knowledge in categories that are related to higher order aspects of writing, such as style, content and the organization of texts, and an increase in knowledge of the writing process, specifically for students in grade 6. Further, we see that students in the experimental condition wrote better texts than their counterparts in the control condition. These findings suggest a relation between increased knowledge and improvement in text quality. We examined this relation in more detail and found that knowledge of the organization of texts and knowledge on the writing process are indeed positively related to text quality. However, our results also show that declarative knowledge of mechanical aspects of writing, such as punctuation, capitals, spelling and grammar was also essential, as this kind of knowledge was also positively related to text quality. Further, we see that knowledge concerning aspects that are related to writing, but not necessarily to text quality (Behavior and Physical), or knowledge concerning aspects that are unrelated to writing (Miscellaneous) did not lead to an improvement in students' writing scores.

The effect size of the intervention on the quality of students' texts was large (ES = 0.72). Although this is in line with research on strategy instruction, this effect is somewhat smaller than is found in several meta-analyses, reporting effect sizes ranging from 0.96 (Koster, Tribushinina, De Jong, & Van den Bergh, 2015) to 1.03 (Graham & Perin, 2007). It must be noted, however, that our intervention was relatively short, compared to the interventions included in the meta-analyses. Closer examination of our results demonstrated that, across all grades, on average, students in the experimental condition scored 10 points higher than their counterparts in the control condition. This progress is quite substantial compared to the average progress per grade, which is 3.35 points. This means that the progress of students in the experimental condition is comparable to three grade levels. Further, students in grade 4 in the experimental condition have higher text quality scores than students in grade 6 in the control condition. Hence, we can conclude that the short intervention in which students were taught a process-based writing strategy was successful in improving the writing performance of students in grade 4 to 6. As we have used a posttest-only design, we cannot rule out the possibility that the writing performance of the students differed between conditions before the start of the intervention. However, as the students in the experimental condition were from eight classes from different schools, it is highly unlikely that students from these classes were all far better writers than students in the control condition.

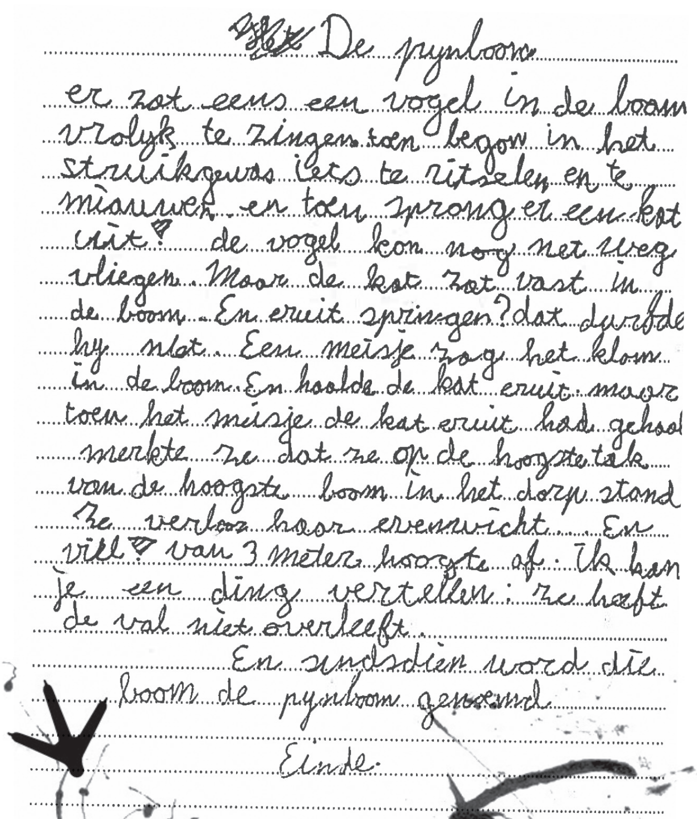
By asking students to give advice on how to get good grades for writing, we also gained insight in classroom practice concerning the teaching of writing, as we can assume that, especially in the control condition, the advice partly reflects the feedback that students normally receive from their teacher in response to their own written products. In line with previous research (Henkens, 2010; Pullens, 2012; Schoonen & De Glopper, 1996: Van der Leeuw, 2007) our results display that the emphasis in writing education mainly concerns the mechanics of writing: during their writing instruction teachers focus mainly on lower order aspects and the written product, instead of providing support to their students concerning the writing process. As was demonstrated by Hamman and colleagues (2000), teachers spend only little of their time on strategy teaching, which is a missed opportunity, as strategies can promote student learning beyond a specific instructional moment. It is however promising that even a short intervention already affects students' knowledge of writing as well as students' writing performance. This shows that when teachers are provided with the requisite materials, they are able to assist their students during the writing process, which has a positive influence on students' knowledge and writing performance.

It should be noted that in this study we only established the short-term effects of the intervention, further research is needed to examine whether these effects last on a longer term. A meta-analysis on intervention studies on self-regulated learning has established that longer interventions are more effective (Dignath & Büttner, 2008). Through a longer period of intensive practice students' strategy use becomes more automated and sophisticated. More opportunities to practice also will promote the transfer of strategies to new situations (Alexander, Graham, & Harris, 1998).

A limitation of the study is that we measured students' knowledge about writing in an indirect way, by having them formulate writing advice to a fictitious peer. We acknowledge the fact that in this way we primarily measured knowledge that was easy to verbalize. It may well be that students have and apply strategic knowledge without consciously paying attention to it. Having students perform a writing task while thinking aloud could provide more insight in how they actually approach a task and to what extent they display strategic behavior during task performance. On the other hand, there were also students who formulated advice that they clearly did not use themselves: e.g., stress the importance of the use of capitals, and not using any in their own letter. These students do have declarative knowledge on these aspects of writing, but still lack the procedural and conditional knowledge to apply this declarative knowledge to a writing task. Further, the design of this study did not allow for direct measurements of conditional knowledge. The intervention was aimed to enlarge students' declarative, procedural and conditional knowledge, as students need conditional knowledge to be able to apply declarative and procedural knowledge to writing tasks. The increased text quality scores of students in the experimental condition suggest that students' conditional knowledge has increased, but we cannot make claims about this. Further research, for instance by using think aloud procedures during writing, could provide more insight in this matter.

Another limitation is that we measured writing performance of students with only one writing task in only one domain of writing. This does not warrant generalization to other writing domains, or to students' writing proficiency in general (Bouwer, Béguin, Sanders, & Van den Bergh, 2015). One task provides an indication of the writing performance of a student in this particular domain: to make any claims about students' writing performance, writing proficiency should be measured with multiple writing tasks and multiple text types.

To conclude, through the combination of a correlational study and an intervention study we have demonstrated that the amount of declarative, procedural, and conditional knowledge that student possess is related to writing proficiency: the more students know about writing, the better the quality of their written texts. Especially knowledge about punctuation and capitals, spelling and grammar, presentation, organization and the writing process contributes to text quality. Traditionally, instruction on punctuation and capitals, spelling and grammar, and presentation form a large part of the writing curriculum, which is reflected by the amount of advice in these specific categories. However, this study shows that in teaching writing attention must be also be devoted to higher order aspects of writing, such as organization and aspects of the writing process. The present study demonstrates that by enhancing students' metacognitive knowledge of writing through teaching strategies and procedures to approach writing tasks leads to large improvements in text quality. Thus, to improve the writing proficiency of elementary students, teachers should focus more on the writing process and offer their students support before and during writing by teaching them writing strategies and when and how these strategies should be applied to attain their writing goals.



~~Het~~ De pynboom

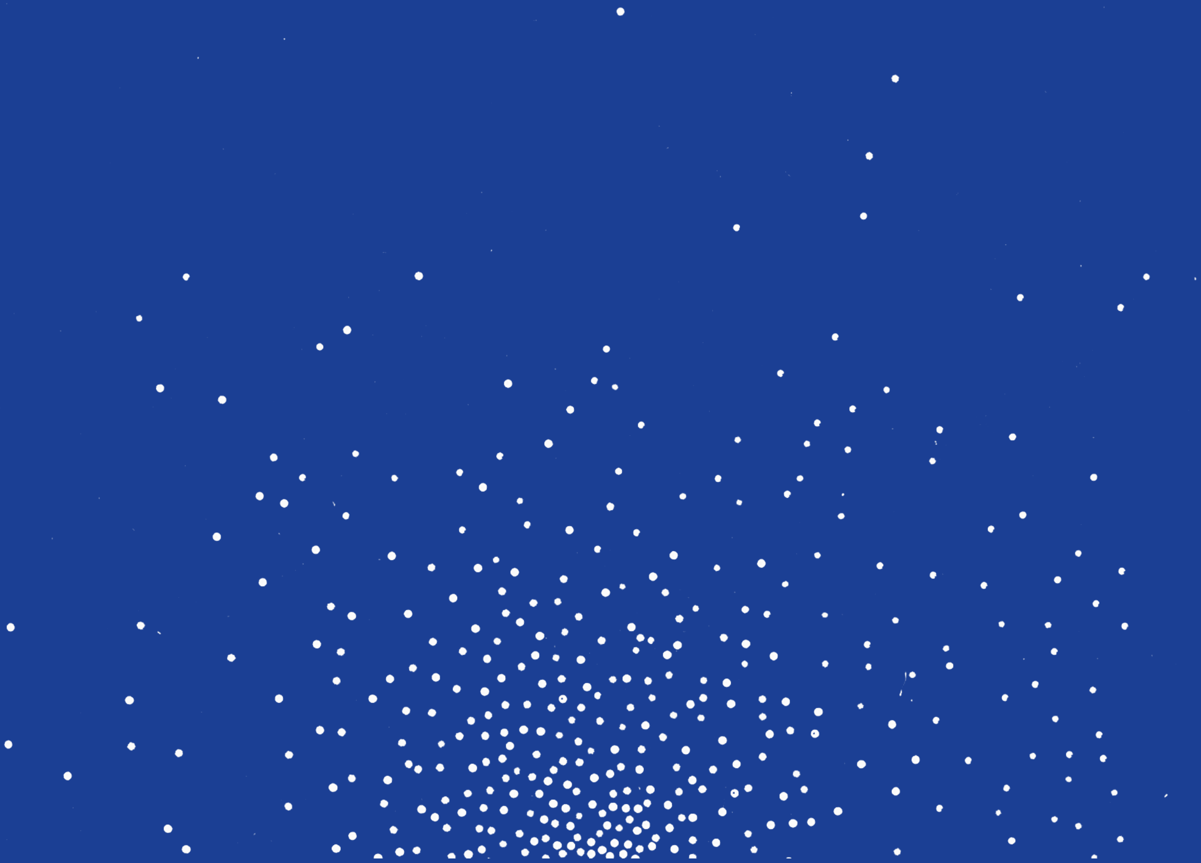
er zat eens een vogel in de boom
vrolyk te zingen toen begon in het
struikgwas iets te ritselen en te
miauwen en toen sprong er een kat
uit! de vogel kon nog net weg
vliegen. Maar de kat zat vast in
de boom. En eruit springen? dat durfde
hy niet. Een meisje zag het klom
in de boom. En haalde de kat eruit maar
toen het meisje de kat eruit had gehaal
merkte ze dat ze op de hoogste tak
van de hoogste boom in het dorp stond
ze verloor haar evenwicht. En
viel! van 3 meter hoogte af. Ik kan
je een ding vertellen: ze heeft
de val niet overleeft.
En sindsdien word die
boom de pynboom genoemd

Einde.

Chapter 7

THINK FIRST! THE EFFECT OF PREWRITING ON THE WRITING PERFORMANCE OF ELEMENTARY STUDENTS

This chapter is based on: Koster, M., Bouwer, R., & Van den Bergh, H. (2016). Think first! The effect of prewriting on the writing performance of elementary students. Manuscript submitted for publication.



This study examined the effect of prewriting strategies on writing performance. During an intervention students ($N=1365$) were taught to generate content and organize content prior to writing. The effects of prewriting were examined in three different grades and with three different types of texts: narrative texts in grade 4, descriptive texts in grade 5, and persuasive texts in grade 6, on three measurement occasions. Results show increased use of pre-writing in all grades, compared to a control group. Findings also indicate a shift in the type of prewriting used after the intervention: less writing drafts, more listing and organizing ideas. The use of prewriting strategies led to higher quality texts, in all grades. Organizing ideas was the most effective type of prewriting (ES = 0.68), followed by listing ideas (ES = 0.43). Least effective was composing a full draft (ES = 0.23). At a delayed posttest measure the effects of the intervention were still visible.

INTRODUCTION

In our rapidly digitalizing society, writing is more important than ever. Communication that used to be oral is being replaced by written communication, such as e-mail and texting. To fully participate in society, it is of vital importance to possess adequate writing skills. Writing is important for the cognitive development of students, and the level of writing performance influences students' academic achievement, school success, academic career and occupational success (Graham & Perin, 2007; Inspectorate of Education, 2012; Langer & Applebee, 1987).

Whereas oral language skills are acquired naturally at home in daily life during the early years of development, writing skills have to be acquired in a school setting through instruction and practice (Inspectorate of Education, 2012; Rijlaarsdam et al., 2012). However, at present, worldwide concerns are raised on the level of students' writing proficiency, in the US as well as in Western Europe (cf. National Center for Education Statistics, 2012; Ofsted, 2012). In the Netherlands, consecutive large Dutch national assessment studies (Krom, Van de Gein, Van der Hoeven, Van der Schoot, Verhelst, Veldhuijzen & Hemker, 2004; Kühlemeier, Van Til, Feenstra, & Hemker, 2013) showed that more than 60% of students in grade 6 are not able to compose texts that sufficiently convey a simple message to a reader. The Dutch Inspectorate for Education found that in the average classroom attention and time devoted to writing are limited, and that the majority of teachers do not succeed in effectively teaching writing (Henkens, 2010). Effective writing instruction is of vital importance to learn to write well, as writing is a very complex process during which many cognitive activities have to be performed simultaneously: generating content, translating ideas into logical and correct sentences whilst taking into account the communicative goal and intended audience of the text, and, at the same time, also paying attention to spelling, grammar, punctuation, word choice, and genre conventions. Especially with beginning writers, this simultaneous activity can lead to cognitive overload which has detrimental effects on text quality (McCutchen, 1996).

Flower and Hayes (1981) were the first researchers to capture the cognitive complexity of writing in a process model. Their model demonstrates how task environment, working memory and long term memory interact during planning, translating, and revising, the different phases of the writing process. Experienced writers regulate these activities with a monitor: this component monitors the progress towards the writing goal, allowing the writer to make adjustments during writing. Further, this monitor regulates switching between the various cognitive activities, which allows for managing a writer's cognitive overload (Flower & Hayes, 1981).

The model of Flower and Hayes, however, reflects the writing process of an expert writer, already capable of regulating his own writing process. The writing process of beginning writers looks quite different (Bereiter and Scardamalia, 1987). According to Bereiter and Scardamalia, beginning writers circumvent cognitive overload by adopting a so-called knowledge-telling strategy. Writers who use this strategy generate an idea, write it down, generate the next idea, write this down, and so on, with the text finished when the writer has run out of ideas. Generally, this approach leads to texts with little structure and coherence (Bereiter & Scardamalia, 1987; Kellogg, 2008). In contrast, more experienced writers adopt a more mature writing strategy, called knowledge-transforming (Bereiter and Scardamalia, 1987). When a writer employs a knowledge-transforming strategy she retrieves ideas from long term memory, evaluates them and transforms them to suit the rhetorical goal of the text. This writing strategy resembles the writing process of experienced writers described in the model of Flower and Hayes (1981).

To improve the writing performance of beginning writers, it is essential to bridge the gap between knowledge-telling and knowledge-transforming. Instruction can help students who largely adhere to a knowledge-telling approach to adopt knowledge transforming procedures (Bereiter, Burtis, & Scardamalia, 1988). To achieve this, instruction should be aimed at developing students' skills to regulate their own writing process and manage their cognitive overload. Teaching students strategies for writing, especially strategies that subdivide the writing process into separate steps (e.g., planning, translating and revising), can be helpful to reduce the cognitive overload students experience during writing, as the number of simultaneously active cognitive processes is reduced (Kellogg, 1988; Zimmerman & Risemberg, 1997). Various meta-analyses have demonstrated that explicit strategy instruction boosts students' writing performance (Graham et al, 2012; Graham & Perin, 2007; Koster et al, 2015). Young and inexperienced writers might especially benefit from prewriting strategies, such as generating content and organize this content before writing the actual text, as during writing they can focus almost exclusively on formulating the text. A large number of studies have reported positive effects of prewriting activities on writing performance, even with students in grade 2 (Saddler, Moran, Graham, and Harris, 2004) and grade 3 (Tracy, Reid, & Graham, 2009), which suggests that students can already benefit from targeted instruction in prewriting strategies from a very young age. Several studies show that engagement in prewriting strategies, such as generating ideas, organizing ideas and planning leads to longer texts, which are better organized and of better quality than texts of students who do not apply prewriting strategies. This was found for elementary students (Brodney, Reeves, & Kazelskis, 1999; Bui, Schumaker, & Deshler, 2006; Chai, 2006; Tracy et al., 2009; Saddler et al., 2004; Zhang & Vukelich, 1998), for secondary students (Chai, 2006; Limpo, Alves, & Fidalgo, 2014; Zhang & Vukelich, 1998), as well as college and university students (Ferrari, Bouffard & Rainville, 1998; Galbraith & Torrance, 2004; Galbraith, Ford, Walker, and Ford, 2005: Kellogg, 1988; Piolat & Roussey, 1996; Rau, 1996).

Limpo et al. (2014) show the importance of explicit instruction for the effective use of prewriting strategies for elementary students. They investigated the effect of prewriting activities of students in grade 4 to 6 and students grade 7 to 9. These students were instructed to use prewriting before writing a story, but they had not received any training or practice in how to use prewriting strategies. The findings of this study show that in contrast to students in grade 7 to 9, for students grade 4 to 6 prewriting was not associated with improvements in text quality (Limpo et al., 2014). Although Bereiter and colleagues (1988) have demonstrated that planning becomes more sophisticated with grade, the results of the study of Limpo and colleagues (2014) cannot be explained by maturation alone. Berninger and colleagues have established that advanced planning skills emerge in the period from grade 4 to 6 (Berninger, Cartwright, Yates, Swanson, & Abbott, 1994; Berninger, Whitaker, Feng, Swanson, & Abbott, 1996; Berninger, Yates, Cartwright, Rutberg, Remy, & Abbott, 1992). However, to support the full development of these skills students need targeted instruction already in the early stages of learning to write. In fact, several studies have demonstrated that younger students can be taught to effectively apply prewriting (cf. Tracy et al., 2009; Saddler et al., 2004).

To optimize instruction in prewriting strategies, it is important to assess the effectiveness of different types of prewriting. Prewriting strategies can roughly be distinguished in three different types: (1) generating ideas: when a student lists ideas without any structure or organization, (2) organizing ideas: when the prewriting displays some form of organization, such as an outline, a scheme or mind map, and (3) writing (full) drafts. Research that compared the effectiveness of different types of prewriting demonstrated that organized and structured drafts were related to higher writing scores, compared to unorganized notes or long composed drafts (Chai, 2006; Piolat & Roussey, 1996). For older students (college and university students) outlining is even more effective (Galbraith & Torrance, 2004; Kellogg, 1988; Rau, 1996). Whereas writing full drafts can be considered a knowledge-telling approach, generating content and subsequently structuring this content before writing can be regarded as knowledge-transforming, which might explain the effectiveness of this specific type of prewriting, besides reducing the cognitive overload during writing (Galbraith et al., 2005; Kellogg, 1988). Thus, the use of effective prewriting strategies might be a key aspect in bridging the gap from knowledge-telling to knowledge-transforming.

The aim of this study is to gain more insight into the specific contribution of prewriting in the improvement of students' writing performance. It has been established that during writing instruction teachers focus primarily on the written product and aspects thereof (writing topic and text type, but also spelling, grammar and punctuation), instead of providing their students support concerning aspects of the writing process, such as promoting prewriting or revising activities (Franssen & Aarnoutse, 2003; Henkens, 2010; Pullens, 2012). To address this issue, we developed Tekster [Texter], a comprehensive teaching program for writing for grade 4 to 6 (Koster, Bouwer, & Van den Bergh, 2014a, 2014b, 2014c). The core of Tekster is a writing strategy that divides the writing process into several separate steps, to reduce students' cognitive overload during writing. Prewriting is an essential part of the strategy, as students are specifically taught to generate content first and organize this content before writing the actual text.

Tekster has proven to be effective in improving students' writing performance as has been shown by two large-scale intervention studies (Bouwer, Koster, & Van den Bergh, 2016a; Koster, Bouwer, & Van den Bergh, 2016b).

In this study we examine whether students use more prewriting due to the intervention, compared to a control group, and we examine whether the intervention leads to changes in the type of prewriting that they use. Further, we investigate whether students who engage in prewriting write higher quality texts than students who do not apply prewriting. Lastly, we examine the effectiveness of the different types of prewriting: is generating and subsequently organizing ideas indeed the most effective type of prewriting, as previous research suggests?

METHOD

Participants

This study is part of a larger research project on the improvement of writing proficiency of students in the upper elementary grades. In total 1365 students from 65 classes from 25 elementary schools participated in the study. The schools were spread all over the Netherlands: 10 schools were located in the northern region, 9 in the middle region, and 6 in the southern region. The participating classes were assigned to two groups. Table 7.1 presents the number of classes and students per grade for each group. There were no differences between the groups in the percentage of female and male students ($\chi^2(2) = .25$, $p < .62$). Students' age ranged from 8 to 13 years, with an average of 10.23 years ($SD = 1.00$).

Table 7.1 Participant information per group and per grade

	Group 1			Group 2		
	N classes	*N* students	% females	*N* classes	*N* students	% females
Grade 4	10	178	49	16	283	48
Grade 5	12	197	49	13	230	50
Grade 6	13	227	49	14	250	52
Total	35	602	49	43	763	50

Note. There were 6 multigrade classes in group 1 and 5 multigrade classes in group 2; therefore, the total number of classes per grade exceeds 65.

Design

The intervention study used a design with switching panels (Shadish, Cook, & Campbell, 2002), with two groups and three measurement occasions, see Table 7.2. In this design, the intervention is implemented in both groups, but at different moments in time. At each measurement occasion students wrote three texts: a narrative, a descriptive, and a persuasive text. During the first period (between the first and the second measurement occasion) the first group worked with the intervention, the second group served as a control group, engaging in their regular writing activities and routines. After the second measurement occasion, the second group started with the intervention, while the first group returned to their regular writing activities and routines. The intervention period lasted four months during which regular classroom teachers delivered one lesson a week (16 lessons in total). For the first group the first measurement occasion served as a pretest, the second as a posttest, and the third as a delayed posttest, which made it possible to test whether the effect of the intervention remained over time. For the second group the first two measurement occasions served as pretests and the third measurement occasion as a posttest.

Table 7.2 Design with switching replications with two groups, three measurement occasions and nine writing tasks (tasks a to i)

	M1	Phase 1	M2	Phase 2	M3
	Tasks	(16 weeks)	Tasks	(16 weeks)	Tasks
Group 1	a, b, c	*Intervention program*	d, e, f	Regular program	g, h, i
Group 2		Regular program		*Intervention program*	

Selection of writing samples

The data of this study are a subsample of the data previously collected in an intervention study in which the effectiveness of the writing program Tekster was tested (Koster, Bouwer, Van den Bergh, 2016b). Every student wrote nine texts in total, but only a subsample of three texts per student was analyzed for this study. As we wanted to get an impression of the effectiveness of prewriting in different grades as well as in different genres, we selected in each grade the texts of one genre. Research has shown that genres differ in complexity, which might affect students' writing process and task performance (cf. Reed, Burton, & Kelly, 1985). Generally is found that persuasive texts are the most complex writing tasks, requiring the most planning (Beauvais, Olive, & Passerault, 2011). Narrative writing tasks are relatively easy, for writers of all proficiency levels (Reed et al., 1985). Reed and colleagues found that narrative writing skills were well developed, to the point of automaticity. The level of difficulty of descriptive writing tasks is comparable to narratives, although they are somewhat more challenging for less proficient writers (Crowhurst & Piche, 1979; Reed et al, 1985). Thus, placed on a continuum according level of difficulty, narratives would be considered easy, descriptives intermediately difficult, and persuasive tasks difficult. We therefore decided to analyze the narrative texts of students in grade 4, the descriptive texts of students in grade 5, and the persuasive texts of students in grade 6, 3680 texts in total. Table 7.3 gives an overview of the number of texts we analyzed, per grade, per group and per measurement occasion.

Writing prompts

Each task contained a writing prompt, with an illustration with relevance to the topic. Each task contained a blank page for prewriting or drafting. Students were free to use this page; they were not specifically instructed to engage in prewriting activities. Appendix G provides examples of writing prompts for each genre. Teachers administered the writing tasks to their students during normal class time, without providing any additional instruction. Students worked individually, without any time restriction.

Table 7.3 Number of texts per grade and per measurement occasion, per group

	Group 1			Group 2		
	M1	M2	M3	M1	M2	M3
	N text	*N* text	*N* text	*N* text	*N* text	*N* text
Grade 4	186	180	185	226	241	231
Grade 5	187	187	176	221	204	210
Grade 6	215	216	181	225	209	200

Writing intervention

The intervention program consisted of a teaching program, Tekster, which included three lesson series of 16 lessons, one for each grade level, compiled in a workbook for students, and an additional teacher manual (Koster, Bouwer, & Van den Bergh, 2014a, 2014b, 2014c). To address the focus as well as the mode of writing instruction, several effective practices were combined in the program, see Koster et al. (2016b) for a more detailed explanation of the program.

The core of the program was an overall writing strategy, in which the writing process was subdivided into three subsequent phases: prewriting, writing and postwriting. To support students in applying the strategy, they were taught a mnemonic representing the steps of the writing process: VOS (fox) for grade 4, DODO (dodo) for grade 5, and EKSTER (magpie) for grade 6. The letters of the acronyms represented the steps in the writing process as follows: VOS (fox) for Verzinnen (generate content), Ordenen (organize), Schrijven (write); DODO (dodo) for Denken (think), Ordenen (organize), Doen (do), Overlezen (read); Ekster (magpie) for Eerst nadenken (think first), Kiezen & ordenen (choose & organize), Schrijven (write), Teruglezen (reread), Evalueren (evaluate), Reviseren (revise).

The lessons of the program were highly structured, see Appendix A for a sample lesson. In every lesson a prewriting phase was included during which students had to generate content for five minutes. To discourage the writing of full draft versions during the prewriting stage, students were instructed to use keywords. Subsequently, students had to organize this content in a scheme which provided the structure of the text. This completed scheme served as the basis for the text-to-be-written. The prewriting phase of all lessons was guided by the teacher. During the first lessons teachers had to model how to generate content and how to use keywords, subsequently they had to model how to organize these keywords in a prestructured scheme. During the next phase they had to model how to compose the text using the completed scheme. To support modeling, teachers were provided a teacher manual. The manual contained a general introduction in which the goal and approach of the program were explicated, as well as the structure of the lessons. For each lesson a detailed lesson plan was included providing an overview of the instruction and activities of the lesson with time planning for each phase of the lesson. These lesson plans described the activities the teacher was expected to execute during the lesson and provide suggestions when to use modeling during instruction. The manual also included a dvd with example videos for all phases of the lessons.

In the lessons students learned to apply the writing strategy to various types of texts, for which authentic writing tasks with various communicative goals and audiences are used. For instance, in each grade they learn to write descriptive texts (e.g., a self-portrait or personal ad), narrative texts (e.g., a story or newspaper article), persuasive texts (e.g., a nomination email for a television program or a flyer for recruiting new members for a club), instructive texts (e.g., a recipe, rules for a game) and personal communication (e.g., a holiday postcard or invitation). The writing assignments were developed in close collaboration with elementary teachers to ensure that the topics would match students' interest and developmental level.

Rating text quality

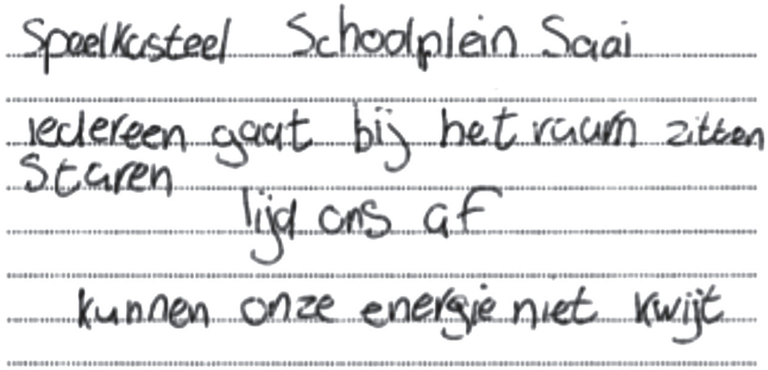
The quality of students' texts was rated by eighteen experienced elementary teachers. Each text was rated by three raters, through the use of overlapping rater teams (Van den Bergh & Eiting, 1989). The texts were anonymized and randomized to ensure that teachers were blind to conditions. The teachers independently scored each text by comparing it to a continuous scale with five benchmark essays, representing the range of writing quality of students in grade 4 to 6. The center position on this scale was an average essay that was assigned an arbitrary score of 100 points, the other essays on the scale were one (115 points) and two (130 points) standard deviations above average, and one (85 points) and two (70 points) standard deviations below average. An example of a rating scale can be found in Appendix K. Raters compared the students' essays to the benchmark essays and administered each essay a score accordingly. Raters were trained in advance on how to apply the scale. The final text quality score of each text was obtained by averaging the scores of three raters. The reliability for each rater was assessed as well as the reliability of each jury. The average reliability of the juries was high, $\rho = .88$, varying from $\rho = .83$ to $\rho = .90$ per task.

Coding prewriting

In contrast to the lessons in the program, students were not prompted to apply prewriting during the writing tasks used in this study. Each writing task contained a full page for prewriting or drafting, which students were free to use. For all writing tasks selected for this study we first coded whether students had applied prewriting or not. This was determined on the basis of the use of the prewriting/drafting page on the writing task. When the prewriting/drafting space was used by a student, this was considered as prewriting. Next was coded, for all tasks in which students applied prewriting, what type of prewriting was applied. We distinguished between four categories: (1) ideas only: when a student lists ideas and content they want to include in their story, without any form of structure or organization (see Figure 7.1 for an example); (2) organized ideas: when the student's prewriting displays some form of organization, such as an outline, a scheme, a mindmap, or arrows or numbers to link related ideas (see Figure 7.2 for an example); (3) draft: when a student has written a version of the text (or a part of it), in full sentences, closely resembling the final version (see Figure 7.3 for an example); (4) other: such as drawings or remarks unrelated to the task.

Coding was done by the first author and 4 trained research assistants. A random subset of 50 texts was coded twice, to establish the interrater reliability. The reliability between raters was very high, Cohen's $\kappa = .92$.

Figure 7.1 Example of listing ideas, without structure or organization.

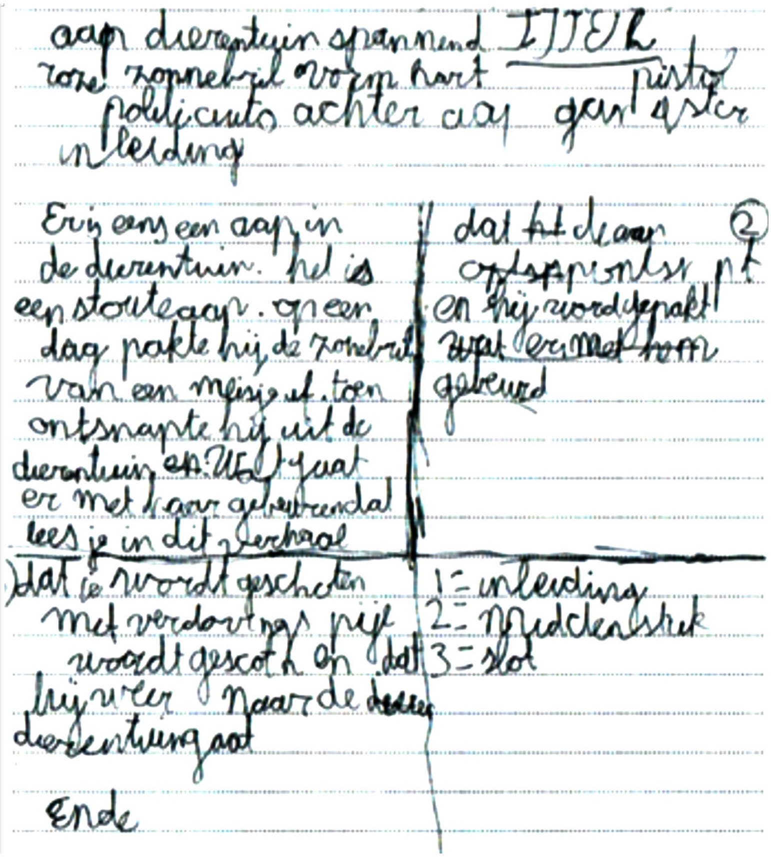


Speelkasteel Schoolplein Saai
iedereen gaat bij het raam zitten
Staren
lijd ons af
kunnen onze energie niet kwijt

Data analysis

First, we investigated whether students used prewriting and whether the use of prewriting changed due to the intervention, separately for grade 4 (narrative writing task), grade 5 (descriptive writing task) and grade 6 (persuasive writing task). As observations are nested within students and students are nested within classes, we used multilevel modeling to analyze the data. To examine whether the probability of prewriting differed over time and/or between groups, we tested, consecutively for each grade, a model in which we included measurement occasion, group and the interaction of measurement occasion and group as

Figure 7.2 Example of organized ideas.



aap dierentuin spannend IJJEL
roze! zonnebril vorm hart pistol
politieauto's achter aap gan' aster
inleiding

Er is eens een aap in
de dierentuin. het is
een stoute aap. op een
dag pakte hij de zonebril
van een meisje af. toen
ontsnapte hij uit de
dierentuin en: Wat wat
er met haar gebeurt endat
lees je in dit verhaal

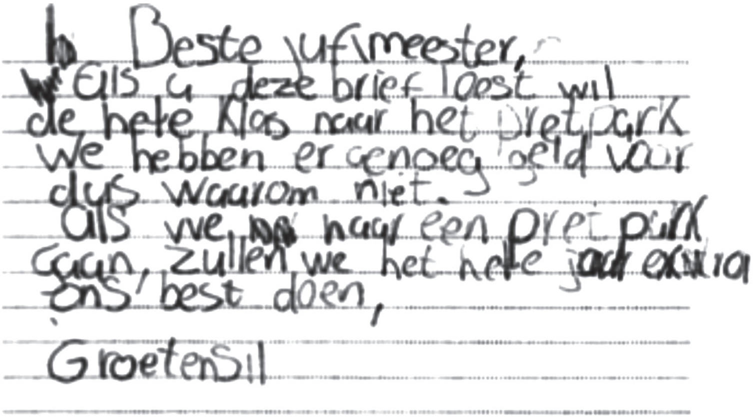
dat het daar ②
opspannendst is
en hij wordt gepakt!
wat er met hem
gebeurd

dat ie wordt geschoten
met verdovings pijl
woordt gescoth en dat
hij weer naar de ~~dieren~~
dierentuin gaat

Ende

1= inleiding
2= middenstuk
3= slot

Figure 7.3 Example of a full draft.



Beste juf\meester,
Als u deze brief leest wil
de hele Klas naar het pretpark
we hebben er genoeg geld voor
dus waarom niet.
Als we naar een pret park
gaan, zullen we het hele jaar extra
ons best doen,

Groeten Sil

explanatory variables. We were mainly interested in interaction effects, as the two groups received the intervention at a different period of time.

Next, we tested whether there were changes over time in the type of prewriting students used. To examine this, we tested separately for each grade and for each type of prewriting (ideas, organized ideas, and draft), a multilevel model with the probability of the type of prewriting as dependent variable and measurement occasion, group and the interaction of measurement occasion and group as explanatory variables.

Subsequently, we examined the effect of (type of) prewriting on text quality. To test this, we fitted three multilevel models to the data, for each grade separately. For each grade, starting point for model fitting was a model with text quality as the dependent variable, and with condition, measurement occasion and the interaction effect of time and measurement occasion as explanatory variables. In the second model, we included prewriting as an explanatory variable in order to test the effect of prewriting on text quality. In the third model, we included different types of prewriting, in order to investigate the effect of the type of prewriting (ideas, organized ideas, and draft) on text quality.

RESULTS

Students' use of prewriting

The results of the multilevel model analyses for all three grades were similar. The effect of time was significant: $F(2, 1243) = 11.07$, $p < .001$ (grade 4), $F(2, 1171) = 7.17$, $p < .001$ (grade 5), and $F(2, 1256) = 7.08$, $p < .001$ (grade 6). There was no significant main effect of group: $F(1, 1243) = 0.28$, $p = .60$ (grade 4), $F(1, 1171) = 1.47$, $p = .22$ (grade 5), and $F(1, 1256) = 1.00$, $p = .32$ (grade 6). The interaction of measurement occasion and group was significant, $F(2, 1243) = 25.49$, $p < .001$ (grade 4), $F(2, 1171) = 13.90$, $p < .001$ (grade 5), and $F(2, 1256) = 13.91$, $p < .001$ (grade 6).

Table 7.4 Mean proportions and standard errors (between parentheses) of use of prewriting per grade and measurement occasion, per group

	Group 1			Group 2		
	M1	M2	M3	M1	M2	M3
Grade 4 (Narrative)	.43 (.13)	.59 (.12)	.40 (.12)	.40 (.11)	.43 (.11)	.81 (.07)
Grade 5 (Descriptive)	.43 (.11)	.58 (.11)	.61 (.11)	.79 (.08)	.52 (.11)	.79 (.08)
Grade 6 (Persuasive)	.32 (.08)	.65 (.08)	.40 (.08)	.55 (.08)	.52 (.09)	.64 (.08)

Table 7.4 gives the estimated means and standard errors of the proportion of prewriting of each measurement occasion for each grade. Comparing students in group 1 and group 2 it can be seen that between the first and second measure-

ment occasion for all grades the use of prewriting increases in the intervention group (Group 1), while in the control group (Group 2) the use of prewriting remains stable or decreases. Between the second and third measurement occasion the effect of the intervention was replicated for Group 2. Thus, after the intervention, the number of students that used prewriting increased: in grade 4 with an average of 27%, in grade 5 with an average of 21%, and in grade 6 with an average of 22.5%. In both groups in all grades more than half of the students applied some form of prewriting after the intervention: ranging from 58% in grade 5 in group 1 (M2), to 81% in grade 4 in group 2 (M3).

Table 7.4 also shows that the use of prewriting for students in grade 5 and 6 in Group 1 at the third measurement occasion was still above pretest level, which indicates that the effect of the intervention remained over time. Remarkably, this effect was not found for grade 4 in Group 1. However, it could be that students stopped using less effective types of prewriting. To meaningfully interpret this result it is essential not only analyze whether students use prewriting or not, but also what type of prewriting they applied and whether this changed due to the intervention.

Type of prewriting

We analyzed 3680 students' texts over three measurement occasions, and almost 60% of the students applied some form of prewriting. Of the students who used prewriting, 50% wrote a full draft, 33% listed ideas, and 15% organized ideas. Less than 1% used another type of prewriting, such as drawings or remarks unrelated to the task, therefore this category was excluded in further analyses.

First, we examined whether there were changes over time in the type of prewriting students used. For all types of prewriting, in all grades we see a significant effect of measurement occasion (ps $< .04$), but we are mainly interested in the inter-action effect of measurement occasion and group, as the two groups received the intervention at different moments in time.

For ideas we found significant interaction effects of measurement occasion and group in grade 5 and 6 ($p = .003$ and $p = .02$ respectively), indicating that there were differences between the groups in the use of this type of prewriting as a result of the intervention. Table 7.5 presents the mean proportions and standard errors of the use of the different types of prewriting, per grade and measurement occasion, per group. It can be seen that between the first and second measurement occasion in grade 5 and 6 the listing of ideas has increased in the intervention group (Group 1), compared to the control group (Group 2). This effect was replicated in Group 2 between the second and third measurement occasion. Although this pattern is the same for grade 4, the interaction effect was not significant ($p = .66$). Lastly, Table 7.5 shows that at the third measurement occasion in Group 1, i.e. the delayed posttest, listing ideas was still above pretest level, which suggests that there is a long-term effect of the intervention.

For organized ideas we found significant interaction effects of measurement occasion and group for all grades (ps $< .04$), which means that in the use of this type of prewriting there were differences between the groups due to the intervention for each grade. In Table 7.5 can be seen that in the intervention group (Group 1) in all grades the use of organized ideas has increased between the first and second measurement occasion, compared to the control group (Group 2). This effect was replicated in Group 2 between the second and third measurement occasion. Further, Table 7.5 shows that at the third measurement occasion in Group 1, i.e. the delayed posttest, the use of organized ideas remained stable, also indicating a long-term effect of the intervention.

Table 7.5 Mean proportions and standard errors (between parentheses) of use of type of prewriting per grade and measurement occasion, per group

	Group 1			Group 2		
	M1	M2	M3	M1	M2	M3
Grade 4 (Narrative)						
ideas	.05 (.02)	.10 (.04)	.18 (.07)	.06 (.03)	.10 (.04)	.26 (.05)
organized ideas	.02 (.01)	.18 (.07)	.14 (.02)	.06 (.02)	.06 (.03)	.40 (.02)
draft	.30 (.09)	.14 (.06)	.10 (.04)	.21 (.07)	.17 (.06)	.15 (.05)
Grade 5 (Descriptive)						
ideas	.09 (.03)	.27 (.07)	.26 (.07)	.09 (.03)	.08 (.03)	.28 (.05)
organized ideas	.03 (.01)	.17 (.03)	.14 (.02)	.02 (.01)	.00 (.00)	.24 (.04)
draft	.21 (.09)	.09 (.04)	.12 (.06)	.51 (.12)	.25 (.09)	.18 (.07)
Grade 6 (Persuasive)						
ideas	.10 (.03)	.30 (.06)	.21 (.60)	.07 (.02)	.10 (.04)	.28 (.06)
organized ideas	.02 (.01)	.11 (.02)	.13 (.03)	.02 (.01)	.01 (.01)	.18 (.03)
draft	.20 (.05)	.20 (.05)	.05 (.05)	.44 (.07)	.33 (.07)	.18 (.05)

For drafts we found significant interaction effects of measurement occasion and group in grade 4 and 5 ($p < .001$ and $p = .03$ respectively). Thus, there were differences between the groups in the use of drafts over time, but they are less straightforward to interpret. Table 7.5 shows that between the first and second measurement occasion in grade 4 draft writing has decreased in the intervention group (Group 1), compared to the control group (Group 2), this effect was less strong in Group 2 between the second and third measurement occasion. In grade 5 between the first and second measurement occasion the decrease in draft writing was stronger in Group 2 than in Group 1, which cannot be attributed to the intervention. In grade 6 there is no significant interaction effect between the first and second measurement occasion ($p = .21$). Although the findings regarding drafting were more dispersed, it can be seen that for all grades in both groups the use of drafts diminished over time. However, this cannot be directly linked to the intervention.

Prewriting and text quality

We examined the effect of prewriting on text quality separately for each grade. For all grades the inclusion of prewriting led to a significant improvement of the model, compared to a model in which we only accounted for the differences in text quality due to group and/or measurement occasion (χ^2 (1) > 22.74; $p < .001$).

Subsequently we added the type of prewriting to the model, this effect was also significant for all grades, (χ^2 (2) > 14.67; p < .001), which indicates that text quality scores differed, depending on the type of prewriting that was used.

Table 7.6 presents the estimates of the regression coefficients of the types of prewriting. It can be seen that all three types of prewriting contributed positively to text quality. Thus, using a prewriting strategy led to a text of a higher quality, for all grades and all text genres, compared to not using a prewriting strategy. As can be seen in Table 7.6, for all grades and genres the most effective type of prewriting was to organize ideas before writing. The magnitude of this effect was estimated by comparing the effect of the type of prewriting to the total amount of variance (Cohen's *d*). The effect size ranged from 0.52 in grade 5 to 0.86 in grade 6. Listing ideas also appeared to be an effective prewriting strategy, with effect sizes ranging from 0.30 (grade 4) to 0.47 (grade 6). Writing full drafts was the least effective type of prewriting, with effect sizes ranging from 0.20 (grade 4) to 0.25 (grade 6).

Table 7.6 Regression coefficients and standard errors for type of prewriting on text quality, per grade

	β	*SE*	p	ES
Grade 4				
ideas	4.68	1.04	< .001	0.39
organized ideas	7.88	1.17	< .001	0.66
draft	2.41	0.89	< .01	0.20
Grade 5				
ideas	5.49	1.04	<.001	0.42
organized ideas	6.76	1.49	<.001	0.52
draft	3.17	1.00	< .01	0.24
Grade 6				
ideas	6.23	0.99	<.001	0.47
organized ideas	11.49	1.56	<.001	0.86
draft	3.34	0.95	<.001	0.25

DISCUSSION

In the present study we examined the effect of prewriting on the writing performance of upper elementary students (grade 4-6). Prewriting constitutes a substantial component of the writing strategies of Tekster, a comprehensive teaching program for writing. In every grade, in every lesson students first have to generate content and organize this content, before starting to write the full version of their text. Through separating content generation and content organization from text composition, cognitive overload during writing is reduced, which should lead to improvements in writing performance. We investigated the effects of the intervention prewriting on the likelihood that students used (different types of) prewriting, and how this affected their writing performance. In this study we included three different grades and three different types of texts: narrative texts of grade 4 students, descriptive texts of grade 5 students, and persuasive texts of grade 6 students, on three different measurement occasions. The results of our study show that, in all grades, students in the intervention group were more likely to use prewriting compared to students in a control group. Furthermore, our findings indicate that due to the intervention, students applied different prewriting strategies than the control group: instead of writing draft versions during the prewriting stage, they preferred listing and organizing ideas as prewriting strategies after the intervention. Using prewriting strategies led to higher quality texts, in all grades. Our findings demonstrate that there are differences in the effectiveness of the various types of prewriting: organize ideas before writing was the most effective type of prewriting, with an average estimated effect size of 0.68, followed by listing ideas (average ES = 0.43). The least effective type of prewriting was composing a full draft (average ES = 0.23). It is promising that the effects of the intervention were visible at a delayed posttest measure: still more students applied prewriting, using more effective types of prewriting (listed ideas and organized ideas) than at pretest measure.

Together, these findings suggest that applying prewriting strategies is a very effective way to improve the writing performance of elementary students, which is in line with previous research (Brodney, Reeves, & Kazelskis, 1999; Bui, Schumaker, & Deshler, 2006; Chai, 2006; Tracy et al., 2009; Saddler et al., 2004; Zhang & Vukelich, 1998). We found that the type of prewriting mattered: generate content and subsequently organizing this content is more effective than writing a draft version of the text, and more effective than generating ideas without structuring them. This corroborates findings from prior studies: for instance Piolat and Roussey (1996) and Chai (2006) also demonstrated that organized ideas are more effective than composed drafts. Writing a full draft from scratch can be considered a knowledge-telling strategy, whereas generating ideas, evaluating these ideas and organizing them before writing reflect a knowledge-transforming strategy. Again, the results from the present study confirm that when students learn to use the latter strategy, they write better organized texts of higher quality. Thus, our findings suggest that targeted instruction in prewriting strategies might be the support that students need to bridge the gap from knowledge-telling to knowledge-transforming. Furthermore, by determining the effectiveness of different types of prewriting, our study provides valuable clues about what the target of instruction should be.

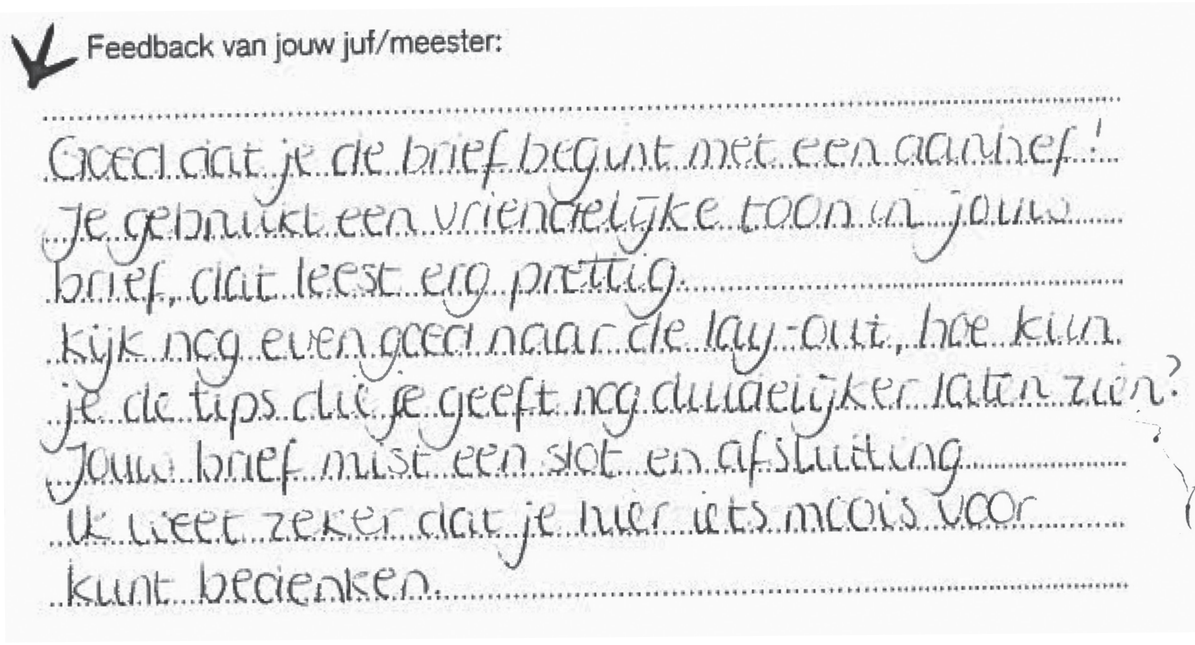
The underlying assumption of the effectiveness of prewriting is that by separating content generation and translation prewriting reduces the writer's cognitive overload during writing. The positive effect of prewriting on text quality seems to confirm this assumption. However, it should be noted that we assessed prewriting solely on the basis of written products: we did not monitor the course of the planning process. Research by Kellogg (1988) suggests that using prewriting strategies particularly alleviates the attentional overload of the writer, as the writer can focus on a single process at the time. Although this leads to improvements in text quality, this seems at the cost of the efficiency of the writing process, as students who use prewriting strategies spent more time on writing (Kellogg, 1988). Torrance and colleagues (2015) also found that using prewriting strategies prolonged the writing process. Concerning the relationship between time spent on planning and text quality, Hayes (1996) demonstrated that in various studies the positive effect of planning on text quality could be attributed entirely to time-on-task: There was no evidence that writers who spent more time on planning were more successful than writers who spent less time on planning. We do not know if this holds true in this study, as we did not monitor the time students spent on writing or on subprocesses of writing. Further research should investigate in more depth the effect of prewriting on the efficiency of the writing process as a whole, and whether prewriting reduces cognitive load by using process measures.

Besides in written form, prewriting can also take place in the writer's head. Research has reported positive effects of mental prewriting, especially on content generation (Kellogg, 1988; Rau, 1996). We cannot exclude the possibility that students, although they did not use the draft paper, did engage in some form of mental prewriting. Process measures, such as thinking-aloud procedures could provide more insight in non-written forms of prewriting.

It is promising that students shifted from applying a less effective prewriting strategy (i.e., writing a full draft) to a more effective strategy (i.e., generating and organizing ideas) and that this change is maintained over time, as shown by the delayed posttest. However, still a substantial number of students did not engage in prewriting activities, despite extensive instruction and practice. Further, although there was a significant increase in the number of students that organized their ideas before writing, it was still a relative small portion of the total number of students. A possible explanation is that the lessons in the program were highly structured and teacher led: students had to generate ideas and organize these ideas in prestructured schemes before starting to write their texts. During the intervention students did not practice independently how to organize their ideas and develop their own schemes. The writing tasks used in this study differed in layout from the writing lessons and offered only a blank page for drafting. Teachers were instructed not to provide their student any additional instruction. In this way we measured the transfer of the strategy to non-prestructured writing tasks. Although positive, the results of this study indicate that there is still room for improvement. This can be achieved in various ways, such as increasing the duration of the intervention, providing more opportunities for students to independently practice with less structured (or unstructured) prewriting activities, or provide regular booster sessions. Future research needs to examine how the program can be further optimized.

In this study we did not examine the content and/or quality of students' writing plans. Thus, we do not know how many ideas students have generated during the prewriting stage, the quality of these ideas and how (and if) they appear in the final text. Analyzing the content of the writing plans could provide more insight in the quality of these plans, and which specific features of writing plans are related to text quality. This could provide valuable clues for teaching.

To summarize, the present study demonstrates that students can benefit greatly from instruction in how to generate and organize ideas before writing. Teaching students prewriting strategies is a promising approach to bridge the gap from knowledge-telling to knowledge-transforming. To promote the use of these strategies beyond the writing lessons it is essential that students recognize the importance of prewriting for their own writing, through explicit instruction and frequent practice.



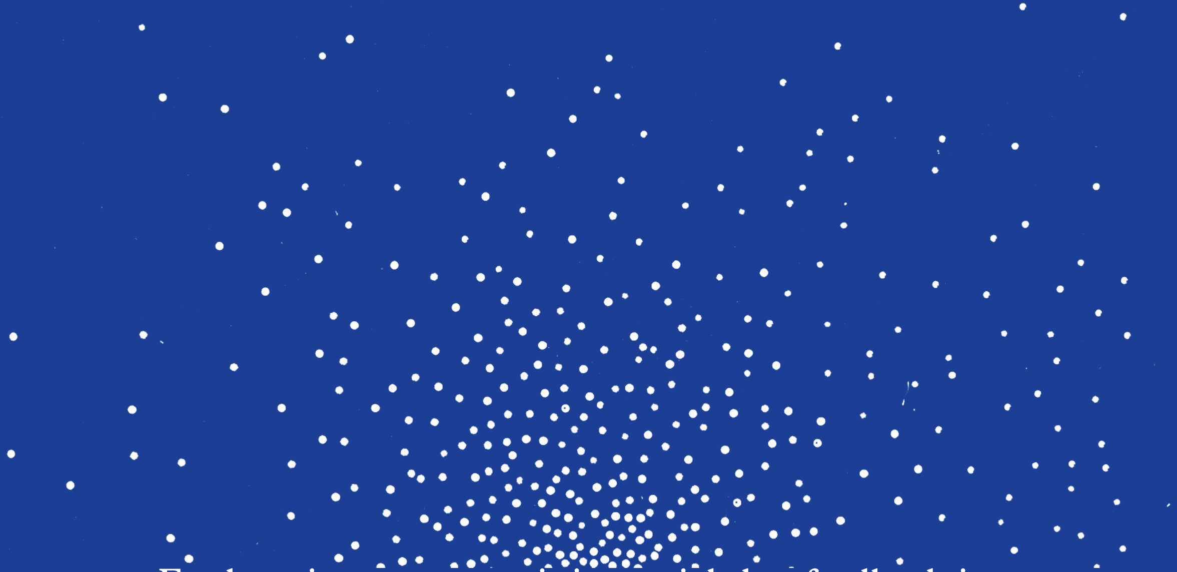
Feedback van jouw juf/meester:

Goed dat je de brief begint met een aanhef!
Je gebruikt een vriendelijke toon in jouw
brief, dat leest erg prettig.
Kijk nog even goed naar de lay-out, hoe kun
je de tips die je geeft nog duidelijker laten zien?
Jouw brief mist een slot en afsluiting
Ik weet zeker dat je hier iets moois voor
kunt bedenken.

Chapter 8

‘WELL DONE, BUT ADD A TITLE!’ FEEDBACK PRACTICES OF ELEMENTARY TEACHERS AND THE RELATIONSHIP WITH TEXT QUALITY

This chapter is based on: Bouwer, R., Koster, M., & Van den Bergh, H. (2016). ‘Well done, but add a title!’ Feedback practices of elementary teachers and the relationship with text quality. Manuscript is submitted for publication.



For learning purposes, it is crucial that feedback is adapted to students' individual needs. The present research investigated how upper-elementary teachers provide feedback and whether they adapt their feedback to students' writing performance. Fourteen experienced upper-elementary teachers from different schools provided feedback on student texts of varying levels of quality. Results show large differences between teachers' feedback practices, indicating that teachers had a preferred style of feedback. On average, teachers provided an equal amount of higher and lower order feedback, primarily focused on negative text features. They rather provided directions than facilitate learning, by textual corrections as well as through comments. Moreover, feedback was hardly related to the quality of the text. When teachers focused primarily on lower order concerns in the text or communicated their feedback with corrections, they did so for poor as well as for high quality texts. The only influence of students' writing performance on teachers' feedback was found for the amount of praise and explicit directions, which was higher for texts of high quality. Taken together, the results from the present study demonstrate that feedback depends more on the style of the teacher than on the quality of the written text.

INTRODUCTION

Feedback is a crucial element in writing instruction (Huot, 2002; Parr & Timperley, 2010). Whereas whole-classroom instructional practices are generally aimed at the average student, the purpose of feedback is to inform students about their individual strengths and weaknesses in order to close the gap between actual and desired performance (Hattie & Timperley, 2007). By providing instruction tailored to students' individual needs, feedback can promote learning for students with different proficiency levels (Wiliam, 2011).

Feedback is especially important for writing, as it helps learning writers to understand whether they adequately communicate a message to a reader (Hillocks, 1982; Zellermayer, 1989). Contrary to oral communication, a writer receives no direct clues or feedback on the communicative effectiveness of their text, as the writer and reader are often separated from each other in time and space. Teachers' responses to students' texts can therefore provide valuable clues about the effects of a text on a reader. Through reader feedback, students experience what constitutes good writing (Brannon & Knoblauch, 1982; Murphy, 2000; Sommers, 1982).

Another reason that feedback is such an effective instructional practice for learning to write is that it supports students in revising their text. Research has shown that revision increases text quality (Crossley & McNamara, 2016). As writing is a cognitively demanding process, students have limited capacity left for monitoring (and enhancing) the quality of their own text during the process of production. This is even more the case for young writers, as they have not yet automatized the lower-level skills such as handwriting and spelling (Fayol, 1999; McCutchen, 2011). But when the first ideas are already translated to text, students can devote all cognitive effort to evaluating whether the text produced so far meets the initial writing goals, and revise when necessary (Chanquoy, 2001). However, simply giving students extra time to revise their text is not sufficient. They need to know and understand what is required in order to revise their text effectively. Feedback can offer guidelines such that students learn how to revise their text themselves (Sadler, 1989).

Although feedback has the potential to improve students' learning and performance, this does not necessarily happen. Kluger and DeNisi (1996) performed a meta-analysis of 131 studies involving feedback interventions across a broad range of areas such as mathematics, reading, memory retention or computer adaptive tests. The results showed that feedback improved student performance in general, but decreased performance in at least one-third of the studies.

Studies of feedback in the field of writing have shown similar results. Recent meta-analyses on effective writing interventions have demonstrated that feedback generally improves elementary students' writing (ESs ranging from 0.77 to 0.88; Graham, Harris, & Hebert, 2011; Graham, McKeown, Kiuhara, & Harris, 2012; Koster, Tribushinina, De Jong, & Van den Bergh, 2015). However, also in the context of writing, feedback is not always as helpful as intended, especially when feedback does not adequately address students' individual needs (Biber, Nekrasova, & Horn, 2011). To understand how feedback can be optimized for learning purposes, the present study investigates teachers' feedback practices in more detail.

Dimensions of effective feedback

Effective feedback answers three questions: where am I going (i.e., what is the goal of writing), how am I going, and where to next (Hattie & Timperley, 2007). When feedback provides specific information that answers these three questions on the level of the text, it is easier for students to revise their texts. However, when feedback is too specific and too detailed, students might become overwhelmed. As a consequence, students either neglect the feedback or copy suggestions without further learning (Brannon & Knoblauch, 1983; Hattie & Timperley, 2007). Therefore, teachers have to direct their attention to the most essential problems in the text, instead of commenting on every feature in the text that can be improved (cf. Biber et al., 2011; Underwood & Tregidgo, 2010). Further, teachers have to frame their feedback in such a way that they do not take over the ownership of the text (e.g., Straub, 1996; Underwood & Tregidgo, 2010). It is therefore important to analyze the content of teachers' feedback (i.e., aspects of the text on which the feedback is focused) as well as its form (i.e., how the feedback is communicated). This results in four dimensions on which teachers' feedback can be evaluated: (1) higher versus lower order, (2) positive versus negative features in the text, (3) directive versus facilitative, and (4) corrections in the text versus comments next to the text.

Higher versus lower order feedback

Feedback that teachers provide on students' texts can be focused on different aspects that underlie text quality (Huot, 1990b; McColly, 1970). For instance, a text can be well written on a higher level regarding content, structure or style, but at the same time it can be poor because of errors in lower order aspects such as grammar, spelling, punctuation or conventions (or vice versa). Previous studies have proposed that feedback should initially be focused on higher order aspects instead of lower order aspects (Underwood & Tregidgo, 2010). The reason for this is that beginning writers do not yet possess the knowledge and skills to reflect on higher order issues by themselves (cf. Chanquoy, 2001). For instance, Sommers (1980) showed that beginning writers, who are less proficient in writing a coherent and understandable text, when asked to revise their text, mainly are concerned with rewording and punctuation. They do not pay attention to higher order concerns yet, as they believe that their first ideas are already well described and need no further modifications. However, when feedback is about higher order concerns, students' attention moves from lower to higher order aspects, which makes higher order revisions more likely. This is in line with findings from other studies showing that higher order feedback increases the time that students spend on revising the content and organization of their text (Covill, 1996; Matsumura, Patthey-Chavez, Valdes & Garnier, 2002) and leads to improvements in students' overall writing performance (Matsumura et al., 2002; Parr & Timperley, 2010).

Whereas higher order feedback is related to successful text revisions, this seems not to be the case when it is mixed with lower order comments. Previous research showed that students who received mixed-level feedback neglected the higher order feedback and mainly focused on lower order feedback to revise local errors in the text (Sommers, 1982; Zamel, 1985). A reason for this persistent focus on lower order concerns is that beginning writers have not yet developed clear goals for writing and experience difficulties in determining what has to be modified in the text and how to change it (Chanquoy, 2001). Feedback that directs students' attention to lower as well as higher order aspects fails to support students in deciding what is important for a good text. As a consequence, students rely on their default process: revising local errors in the text. Hence, to support students in revising higher order aspects, teachers have to prioritize higher order feedback over lower order issues. Underwood and Tregidgo (2010) propose that feedback should be directed on lower level aspects only when a student is able to produce a text with a sufficient level of communicativeness.

Whereas research emphasized the importance of higher order feedback, two large studies on feedback practices of university teachers have shown that teachers focus primarily on lower order aspects (Connors & Lunsford, 1993; Stern & Solomon, 2006). In line with these findings, Clare, Valdés and Patthey-Chavez (2000) examined the feedback of 11 elementary teachers and showed that only 14% of the students received higher order feedback. The elementary teachers who participated in the study of Parr and Timperley (2010) provided a greater percentage of higher-order feedback, but this was still outnumbered by the proportion of lower order feedback.

How can we explain this mismatch between research and practice? One explanation might be that teachers are used to focus on grammar and spelling issues in their language curriculum, and therefore automatically also focus on these aspects when evaluating students' writing. This might especially be the case for elementary teachers, as students in the lower grades have only just mastered the lower-level skills of writing. Further, it is much easier and quicker to respond on lower order issues, as these are either right or wrong. This is different for higher order issues, for which evaluation is more subjective. Another explanation for the large proportion of lower order feedback might be that in a text the occurrence of language-related errors is much higher than the occurrence of mistakes in structure or content. After all, whereas every word can potentially include at least one language-related mistake, errors with regard to content or structure are related to larger units, such as sentences or paragraphs.

To understand why teachers devote such a high proportion of their feedback to lower order aspects, it is essential to first examine whether teachers are focused on lower order aspects in every student text, or whether this depends on students' writing performance. Effective feedback is not about whether the feedback focuses on higher or lower order aspects of the text, but whether it reflects the main problem of the text. For instance, if a text does not fulfill its communicative goal because of missing information or an unclear line of thought, feedback should be focused on these aspects before directing students' attention to lower order aspects. Therefore, when examining the content of teachers' feedback, it is important to take students' writing performance into account. However, previous studies measured the amount of higher versus lower order feedback only on a general level. It is still unclear whether texts of different quality levels receive different kinds of feedback and, hence, whether feedback is tailored to students' individual needs.

Feedback on positive versus negative features

The second dimension of feedback is whether it focuses on text features that already meet the standard (i.e., positive features) or on text features that still need improvement (i.e., negative features). Previous research suggests that feedback on negative features increases students' revising behavior as it challenges them to set higher goals for future performance. The same pressure for further improvement is not necessary when feedback stresses positive features of a performance (Podsakoff & Fahr, 1989). This might be one of the reasons for teachers to focus primarily on negative features in the text, as was demonstrated in two large-scale studies on teachers' feedback practices (Connors & Lunsford, 1993; Stern & Salomon, 2006).

The beneficial effects of feedback on negative features are, however, not as strong for low-ability students as for high-ability students. Podsakoff and Fahr (1989) explain this by students' self-efficacy: low-ability students do generally not believe that they are able to adequately apply the feedback they receive, which holds them back from using it. Feedback on positive features can increase feelings of self-efficacy as it enables students to reinforce appropriate writing behavior (Hyland & Hyland, 2001). Hillocks (1986) analyzed multiple studies on the effect of positive comments and showed that students who received such praise, on average, wrote longer texts and developed more positive attitudes about themselves as writers and about writing as an activity.

However, praise does not always motivate students to write more. For instance, when teachers praise students' efforts in a more or less general way, it moves their attention away from the text and towards themselves as a person. When students' attention is not longer directed at the text, it is less likely that something in the text will be improved (Hattie & Timperley, 2007; Kluger & DeNisi, 1996). Moreover, when the text is already of high quality, there is only little room for further improvement. This does not mean, however, that a student cannot learn from receiving positive feedback. Underwood and Tregidgo (2010) therefore stress that positive feedback has to be linked to specific features of the text in order to encourage students' writing.

Although teachers are aware of the importance of providing positive feedback (Peterson & McClay, 2010; Van den Bergh, Ros, & Beijaard, 2013), they seem to use positive remarks to mitigate criticism, rather than informing students about their writing (Hyland & Hyland, 2001). For instance, Smith (1997) showed that most teachers start their feedback with a positive remark, but then quickly move to suggestions for improvement, sometimes even within one sentence. Further, teachers' positive remarks are often too generic, such that they are easily applicable to every student. Peterson and McClay (2010) interviewed elementary teachers and revealed that they sometimes even wrote 'very good' on texts of low quality. This suggests that teachers struggle with praising specific features of students' writing. Yet, more research is needed to understand whether teachers provide feedback on positive features in the text, and whether this depends on the quality of the text.

Directive versus facilitative feedback

The third feedback dimension evaluates whether teachers' feedback is aimed at directing students how to move forward or at facilitating students' learning (cf. Biber et al., 2011; Black & William, 1998). Directive feedback explicitly tells students what needs to be revised and how. Facilitative feedback, on the other hand, leaves the control of the revision process with the student by providing only reader-based responses such as reflections, suggestions or clarification questions that students can use when revising their text. Brannon and Knoblauch (1982) argued that when teachers leave the control with the student, students reach a deeper understanding of the revision process, as they learn how to change the text to meet communicative purposes instead of meeting teachers' expectations and standards.

Thus, facilitative feedback is generally regarded as more effective for learning to write than directive feedback. Research has shown, however, that not all students are equally capable to understand and use facilitative feedback. For instance, Clare et al. (2000) showed that that in elementary grades only half of the students were able to handle feedback that contained questions or requested for clarifications. This percentage was higher for students in secondary education, with generally more developed writing skills (Clare et al., 2000). These results seems to suggest that the ability to use facilitative feedback depends on students' proficiency level (Clare et al., 2000; Murphy, 2000; Shute, 2008; Ziv, 1984). This is in line with Shute (2008), who proposes that less proficient students need more explicit and specific instructions to move forward. Ziv (1981) concluded the same: beginning writers need explicit directions from their teachers about how to strengthen or reorganize ideas in the text. Hence, these studies emphasize that it is important for teachers to adjust the extent of control and directions in their feedback according to students' level of proficiency.

Error corrections versus comments

The fourth dimension is related to how teachers provide feedback, which can either be with corrections in the text or with comments next to the text. Previous research revealed that students do not learn how to effectively regulate their own writing when teachers provide their feedback in the form of corrections in the text (Straub, 1996): They can adopt suggested corrections without having to engage in deep learning processes. Although such corrections might lead to improvements in the revised text, they do not transfer to other writing tasks, as students do not learn the underlying reason for the suggested corrections (Hattie & Timperley, 2007). This became evident in a case study of Ziv (1984) in which four inexperienced college writers were followed during a yearlong writing course. The results showed that students accepted corrections that were inserted in their texts, but without any understanding why they had been made – they kept making the same mistakes over and over again. In contrast, students who receive feedback in the form of comments show more progress in writing and believe that writing is more important and enjoyable than students who receive corrections (Semke, 1984). Hence, to engage students in revising their own text, teachers should leave the ownership of the text with the student and provide suggestions for improvement in end comments rather than providing corrections in the text.

Aim of the present study

Whereas previous research emphasized over and over again that feedback has to be adapted to students' individual needs, there is hardly any research that relates feedback to students' proficiency level. Whereas previous studies showed that teachers provide feedback primarily on lower order concerns and on features that need further improvement, it is still unclear whether the focus of feedback varies depending on characteristics of the text. The same holds for the dimensions of feedback that evaluate how much control a teacher exerts with the feedback over the text. It is advised that teachers provide suggestions to students on how to improve their text instead of correcting all errors in the text. However, it is still unknown whether the directiveness of feedback depends on students' writing performance. It is especially important to understand whether elementary teachers differentiate their feedback, as beginning writers do not yet have the capacity to regulate their own writing. A recent study has shown, however, that only a minority of teachers in elementary grades believe that feedback should be tuned to individual students (Van den Bergh, Ros, & Beijaard, 2013). Therefore, the central question in this study is whether elementary teachers' feedback depends on students' writing performance: do teachers provide different kinds of feedback depending on the quality of the text?

METHOD

Participants

Fourteen elementary school teachers from 14 randomly selected elementary schools in the Netherlands participated in this study. Twelve of the fourteen teachers were female, and their mean age was 44 years (*SD* = 13.78). On average, they had 15.29 years of experience (*SD* = 13.47; ranging from 2 to 47 years). All teachers were experienced in teaching in the upper elementary grades. They received a financial compensation for their participation.

Material and procedure

The texts on which teachers had to provide feedback were selected from an earlier research project on student's writing performance in grade 6 in which 67 students wrote texts in three different genres: narratives, formal letters, and argumentative essays (Pullens, Den Ouden, Herrlitz, & Van den Bergh, 2012). For narratives, students were asked to write a story about a personal experience in which he or she was being caught. The formal letters consisted of persuasive letters to a fictional company about the collection of toys at a supermarket. Argumentative essays were about the pros and cons of a candy ban for children. The handwritten texts were retyped to control for the effect of handwriting on text quality ratings. The retyped texts resembled the handwritten texts precisely, that is, all errors concerning spelling, grammar, punctuation or capitals were copied, as well as all modifications made by the student. The quality of each text was rated holistically by juries of three raters (for more information, see Pullens et al., 2012). The ratings were satisfactory, with jury reliabilities (ρ) ranging from .73 to .83 per task ($\rho_{average} = .77$).

For each of the three genres five texts were selected based upon their scores for text quality. One text was of average quality, two texts were below average and two texts were above average. The below average texts included one text that was one standard deviation below average (i.e., 16% of the texts were of less quality), and one text that was two standard deviations below average (i.e., only 2.4% of the texts were of less quality). The above average texts included one text that was one standard deviation above average (i.e., that was better than 84% of the texts), and one text that was two standard deviations above average (i.e., better than 97.5% of the texts). This selection procedure resulted in a total of fifteen student texts, representing the range of text quality in upper elementary grades for three writing genres.

Teachers received the fifteen student texts, including prompts, with the instruction to "provide written feedback on these texts as if these were from your own students". After providing feedback, teachers returned the texts plus feedback by mail. In total, there were 14 teachers who provided feedback on 5 texts in 3 genres, resulting in 210 texts with teachers' feedback.

Feedback coding

First, the feedback was split up in segments, such that each feedback segment contained only one single point of feedback (Duijnhouwer, Prins, & Stokking, 2010). When the feedback switched to another aspect (e.g., from punctuation to content), or when it switched from something positive to something negative in the text (or vice versa), the feedback was placed in a new segment. The total number of feedback segments for each text was used as a measure of the amount of feedback.

Second, each feedback segment was coded on four different dimensions using a coding protocol, see Table 8.1. For the first dimension it was evaluated whether a feedback segment was directed on a higher versus lower aspect of the text. For this dimension, we also evaluated the specific object to which the feedback was directed. Feedback on the 'style', 'structure', or 'content' of the text was considered as higher order feedback, whereas feedback regarding 'punctuation', 'spelling', 'grammar', or 'conventions' was considered as lower order feedback. For the second dimension it was evaluated whether the feedback segment addressed a positive or negative text feature. For the third dimension it was evaluated whether the feedback directed the student to a solution or whether it facilitated student's own revision process. For the final dimension it was evaluated whether the feedback was provided with a correction in the text or whether it was communicated with a comment next to the text. A second rater coded all comments of a random selection of 10 percent of the texts. The reliability between raters over all categories was acceptable, Cohen's $\kappa = .72$ (ranging from .54 to .93, depending on the category).

Design and analyses

Feedback segments are nested within teachers ($N = 14$) and within texts ($N = 15$). As dimensions of feedback are likely to vary as a function of the teacher and the quality of a text, they are analyzed by multilevel modeling, allowing for variance due to teachers and texts. For each of the four feedback dimensions we applied two multilevel models. In the first model, the null model, the average probability on a specific type of feedback was estimated, given a random text and a random teacher. The parameter estimates of such models are in logits[1], which are a nonlinear transformation of the probabilities. To enhance interpretation the logits are transformed back to probabilities of occurrence[2]. In the second model, text quality was added as a fixed variable, in order to analyze whether teachers adapted their feedback to the quality of the text.

To interpret and illustrate the quantitative findings, we will provide four feedback examples of two different teachers. These examples were selected based on maximized differences in the way teachers provided feedback. For each teacher we selected two examples of feedback, one for a text of very poor quality and one for a text of very high quality. Based upon the comparison of these feedback examples it is possible to examine how the four dimensions of feedback interact with each other.

RESULTS

Teachers provided a total of 1534 feedback segments on 210 student texts, resulting in an average of 7.30 segments ($SE = 1.36$) per text. This amount varied largely between teachers ($S^2 = 20.75$) and texts ($S^2 = 4.88$): some teachers provided more feedback than other teachers (80% CI [1.47, 13.13]) and some texts received more feedback than other texts (80% CI [4.47, 10.13]). In total 3.5% of the feedback ($N = 53$) was related to an overall judgment about the writer's performance

[1] $\ln \left[\frac{\mathrm{P(ij)}}{\mathrm{1\text{-}P(ij)}}\right]$ [2] $P = \frac{1}{[1+e^{\text{-logit(estimate)}}]}$

Table 8.1 Protocol for coding teachers' feedback

Dimension	Description	Code	Example	Cohen's κ (*SE*)
Higher versus lower order feedback	Is the feedback aimed at higher order or lower order aspects in the text?	1: lower order 2: higher order	1: Add punctuation 2: What would be a good solution for this?	.71 (.06)
	To what specific aspect of the text is the feedback focused?	*Lower order:* 1: punctuation/ capitalization 2: spelling 3: grammar 4: lay-out/conventions *Higher order:* 5: style/language use 6: structure/organization 7: content	1: dear Mr. has to be Dear Mr. 2: contener has to be container 3: I looked at the window > I saw through… 4: Add a title 5: Don't use too much 'and then' in a sentence 6: Your text structure is clear 7: What do you mean with this?	.59 (.04)
Feedback on positive versus negative features	Is it about something positive in the text or about something that needs improvement?	1: negative 2: positive	1: But what is your opinion? 2: Good title!	.93 (.05)
Directive versus facilitative feedback	Is it directing students to a solution, or is it facilitating students' learning by providing hints, explanations, questions or reader responses?	1: directive 2: facilitative	1: The salutation is missing, add it. 2: I'm confused by the past tense that you use here	.54 (.08)
Corrections versus comments	Is it communicated with markings or corrections in the text or is it described with comments next to the text?	1: correction 2: comment	1: spen > spend 2: The letter could be a little more elaborated	.84 (.04)

(e.g., “well done”) or to the underlying writing process (e.g., “good luck with rewriting your text”). The majority of feedback segments ($N = 1481$) were related to the written product.

Higher versus lower order feedback

The results in Table 8.2 show that the probability of feedback on higher order aspects ($P = 0.47$) was not significantly different from the probability on lower order aspects ($P = 0.53$, Wald $z = -0.49$, $p = .63$), indicating that teachers on average provided an equal amount of feedback on higher and lower order issues. They responded mostly to the content of the text (30%). Lower order feedback primarily concerned issues in punctuation/capitalization (23%) or conventions (15%). Only a limited amount of feedback concerned the style (9%) or structure of the text (3%), or grammatical (9%) and spelling issues (11%). Results further show that the probability of higher versus lower order feedback differed largely between teachers ($S^2 = 0.30$). Hence, an 80% confidence interval for differences between teachers in the proportion of higher order feedback varies from 0.27 to 0.69, indicating that some teachers devoted only a quarter of their feedback to higher order aspects, whereas other teachers devoted at least two-thirds of their feedback to higher order issues. There were also large differences between texts ($S^2 = 0.39$). Hence, the 80% confidence interval for differences between texts indicates that the proportion of higher order feedback on a random text varies from 0.24 to 0.72.

Table 8.2 Estimates of (logistic) multilevel models for different feedback dimensions

		Logit (SE)		
Dimension	**Proportion**	**Intercept**	**$S^2_{teachers}$**	**S^2_{texts}**
Higher order feedback (versus lower order feedback)	.47	-0.11 (0.23)	0.30[a] (0.14)	0.39[a] (0.18)
Feedback on positive features (versus negative features)	.09	-2.37[c] (0.36)	0.95[a] (0.43)	0.74[a] (0.36)
Facilitative feedback (versus directive feedback)	.31	-0.82[b] (0.26)	0.82[a] (0.36)	0.04 (0.05)
Comments (versus error corrections)	.64	0.56 (0.48)	2.90[a] (1.32)	0.18 (0.10)

Note. [a] $p < .05$, [b] $p < .01$, [c] $p < .001$

Feedback on positive versus negative features

Table 8.2 further shows that the proportion of feedback on positive features of the text was only .09, which was significantly lower than the proportion of feedback on problems in the text (.91, Wald z = -6.54, $p < .001$). Hence, teachers were more likely to comment on features that needed improvement than focusing on something the student was doing well. There were large differences between teachers in the likelihood of providing positive feedback ($S^2 = 0.95$). The 80% confidence interval for differences between teachers on the proportion of positive feedback indicates that some teachers hardly responded to positive text features, whereas other teachers devoted one-third of their feedback to positive features (80% CI [0.02, 0.32]). There were also differences between texts ($S^2 = 0.74$), indicating that the amount of positive feedback on a random text varied from 2 to 28% (80% CI [0.02, 0.28]).

Directive versus facilitative feedback

Table 8.2 shows that the probability of facilitative feedback (0.31) was significantly lower than the probability of directive feedback (.69; Wald $z = - 3.16$, $p < .01$). This varied largely between teachers ($S^2 = 0.82$): some teachers hardly provided facilitative feedback and focused on directing students, whereas other teachers provided facilitative feedback in two-thirds of the cases (80% CI [0.09, 0.66]). There were no significant differences between texts.

Corrections versus comments

Further, Table 8.2 shows that the probability of comments (.64) was not significantly higher than the probability of error marking or correction (.36, Wald z = 1.17, $p = .24$). This indicates that feedback is communicated not only through comments in the margin or at the end of the text but also with markings or corrections in the text. The likelihood that teachers responded with comments differed largely between teachers ($S^2 = 2.90$), indicating that some teachers hardly responded with comments next to the text, whereas other teachers provided feedback only with comments (80% CI [0.10, 0.97]). There were no significant differences between texts.

Effect of text quality

To examine whether teachers adapted characteristics of their feedback to individual needs of the student, we included text quality as an explanatory variable in the analyses. With regards to the content of the feedback, results show that there was a significant linear effect of text quality only for the proportion of positive feedback (Wald z = 4.93, $p < .001$)[3] , not for the proportion of higher order feedback (Wald z = -0.77, $p = .44$). For the dimension positive versus negative feedback, text quality explained 76% of the variance related to texts, as the variance between texts decreased from 0.74 to 0.18. Table 8.3 presents the parameter estimates (in logits) and shows that the intercept for positive feedback is -2.42 logit ($SE = 0.31$), which is the natural logarithm of the odds that a teacher provides positive feedback on a text of average quality. For every standard deviation increase in text quality, the natural logarithm of the odds increased with .57 ($SE = 0.12$).

[3] It was also tested whether the relation deviated from linearity, which was not the case for all four dimensions.

Table 8.3 Estimates of logistic multilevel models for each feedback dimension when text quality is included as a fixed variable

		Logit (SE)		
Dimension	**Intercept**	**TQ**	**$S^2_{teachers}$**	**S^2_{texts}**
Higher order feedback (versus lower order feedback)	-0.11 (0.23)	-0.10 (0.12)	0.30[a] (0.14)	0.41[a] (0.19)
Feedback on positive features (versus negative features)	-2.42[c] (0.31)	0.57[c] (0.12)	0.94[a] (0.43)	0.18 (0.12)
Facilitative feedback (versus directive feedback)	-0.84[b] (0.36)	-0.16[b] (0.05)	0.84[b] (0.36)	0.00 (0.03)
Comments (versus error corrections)	0.56 (0.48)	-0.16 (0.08)	2.93[a] (1.33)	0.15 (0.08)

Note. [a] $p < .05$, [b] $p < .01$, [c] $p < .001$

With regards to the form of teachers' feedback, results show that there was a significant linear effect of text quality for the proportion of facilitative versus directive feedback (Wald z = -3.15, $p < .01$), but not for the proportion of comments versus corrections (Wald z = -1.92, p = .06). Table 8.3 shows that the intercept for facilitative feedback is -0.84 logit (SE = 0.26), which decreases with 0.16 (SE = 0.05) for every standard deviation increase in text quality.

Figure 8.1 offers a more readily interpretation of the data, and expresses the relationship between text quality and the proportion of occurrence of each feedback dimension. It shows that (a) the probability of higher order feedback was the same for texts of poor and high quality, and (b) that teachers were more likely to provide positive feedback on texts of higher quality, as the proportion of positive feedback increases from .14 for texts of very poor quality to .61 for texts of very high quality. It further shows that (c) the probability of comments remains the same for texts of poor and high quality, and (d) that teachers are more likely to provide directive versus facilitative feedback, which is even more so for texts of higher quality. The proportion of directive feedback increased from .73 to .84, whereas the proportion of facilitative feedback decreased from .27 to .16.

Examples of teacher feedback

The previous analyses separated characteristics of feedback in four different dimensions. However, to fully understand what a teacher communicates to a student based upon the written text, feedback should also be evaluated as a whole. Figure 8.2a shows two examples of feedback from two teachers on a text of very poor quality. Figure 8.2b shows two feedback examples from the same teachers on a text of very high quality. The teachers are selected based upon their differences: teacher 1 provided feedback primarily next to the text, whereas teacher 2 provided feedback primarily in the text. These examples illustrate how the four feedback dimensions interact for different teachers and for different texts.

Figure 8.1 **Effects of text quality on the estimated probability of (a) higher versus lower order concerns (hoc/loc), (b) feedback on positive versus negative features (pos/neg), (c) directive versus facilitative feedback (dir/fac), (d) comments versus corrections (com/cor).**

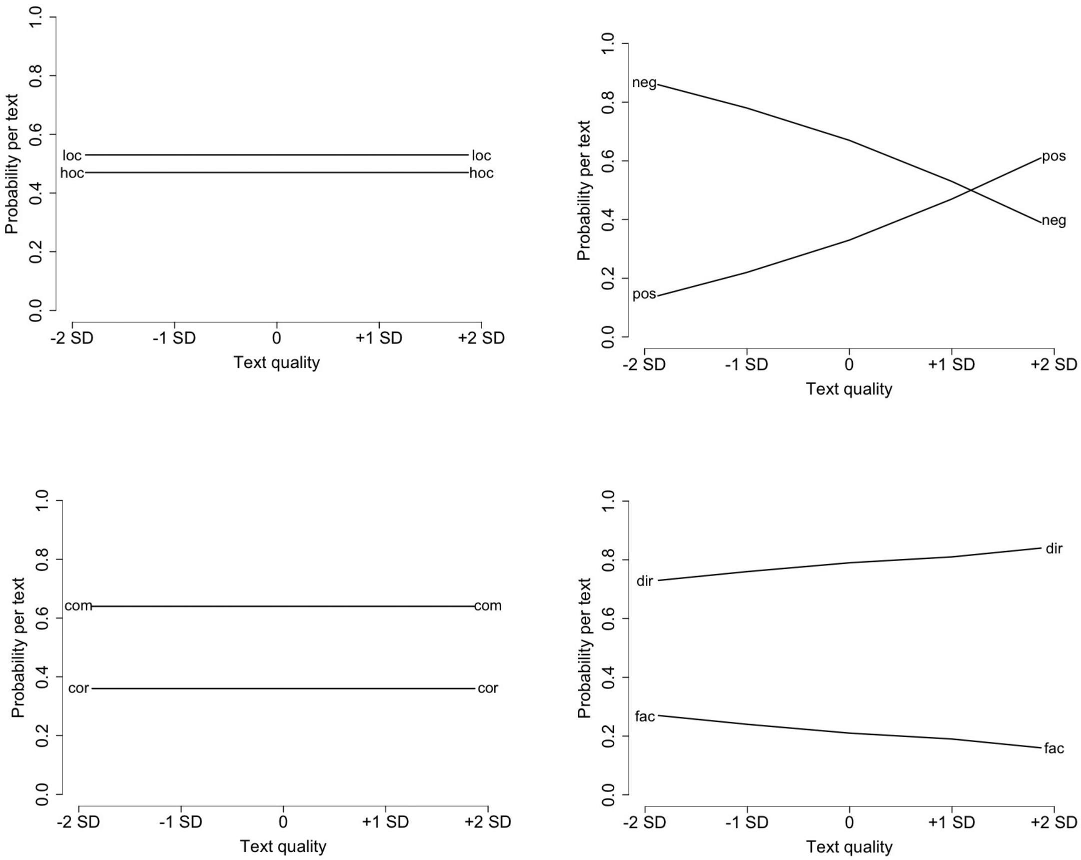


Figure 8.2a shows that both teachers provided feedback on the level of the text, for which they commented on higher as well as on lower order aspects. Further, they were focused only on negative features in the text; neither of them mentioned something the student already did well. The most striking difference between the two teachers is the amount of feedback they provided and how they communicated this to the student. With regards to the amount of feedback, the examples show that teacher 1 provided only 3 points of feedback, whereas teacher 2 provided feedback on every aspect in the text that needed improvement. Although the feedback of teacher 1 does not bridge the complete gap between current and desired performance, it guides the student to a better performance with one step at a time. In contrast, the feedback of teacher 2 is much more detailed and offers a thorough evaluation of the gap between current

Figure 8.2*a* Two examples of teacher feedback (in italics) for a very poor quality text.

TEACHER 1	TEACHER 2
	To the Supercoop firm *Place, date, 2008* *Dear sir/madam,*
i would like to receive the smurfs campaign before April 20th after closing time ~~already please there~~ because I spent 110,34 on groceries and the smurfs ran out. 15/4/2008	i would like to receive the Smurfs campaign before April 20th after closing time ~~already please there~~ *(why?)* because *(what are you trying to say here?)* I spent 110,34 euro's on groceries and the smurfs ran out. ~~15/4/2008~~ → o*n April 15th… (write out)*
First read the assignment again: You want to receive the Smurfs before the closing date, but why? I miss the salutation above the letter as well as the sender. You've written your name and address on the envelope instead of at the bottom of the letter.	*Lay-out: concluding sentence is missing* *name, signature is missing, note own address at bottom of letter* *Language: sentence construction inadequate: I v. … ., because …. (main clause + subordinate clause) use of words inadequate, too little information in the letter* *use capitals for address! is missing in sentences* *use punctuation is inadequate* *Structure: introduction: inadequate, information is missing: father – April 15th packages ran out at supermarket Supercoop in town – receipt* *key point: receipt with stamp – send – please receive package with Smurfs! – campaign runs until April 20th…* *Conclusion: express gratitude – you would like the golden Smurf!!!* *Extra: rewrite!* *Take note: clear structure conclusion-core-intr and determine correct content!*

Figure 8.2*b* Two examples of teacher feedback (in italics) for a very high quality text.

TEACHER 1

Halsteren 15-04-08

Dear Supercoop firm.
My father does his grocery shopping at supermarket Supercoop every week. That's why I already have a whole collection of Smurfs. It's just that I'm still missing 2: Brainy Smurf and Papa Smurf. My father went grocery shopping again today and found out that the Smurfs have run out: I would find it such a shame if that meant I couldn't complete my collection. That's why I would like to ask you if I could still receive Smurfs with the stamp on the shopping receipt. The receipt is in the envelope. If the idea could proceed this is my address: Rode Schouw 47 4661 ve Halsteren

Nice letter!
Just a few tips: you could have mentioned that you would like to receive the Smurfs befóre the closing date of the campaign. Also put your name at the bottom of the letter.

TEACHER 2

take note: salutation – date – to whom
Halsteren, April 15, 2008

To the Supercoop firm

← Halsteren 15-04-08

Dear ~~*Supercoop firm*~~ *sir/madam, (blank line)*
My father does his grocery shopping at supermarket Supercoop every week. That's why I already have a whole collection of Smurfs. It's just that I'm still missing ~~2~~ *two*: Brainy Smurf and Papa Smurf. My father went grocery shopping again today *(which date?)* and found out that the Smurfs have run out: *(. or ;)* I would find it such a shame if that meant I couldn't complete my collection. + *of course I'm hoping for the Golden Smurf!* That's why I would like to ask you if I could still receive *a package of* Smurfs *befóre* April 20th with the stamp on the shopping receipt. The receipt is *enclosed* in the envelope. *I hope this is possible before April 20th*. ~~If the idea could proceed this is my address:~~ ~~*My address is as follows:*~~ Rode Schouw 47
4661 ve
Halsteren

Closing: Sincerely, Name
Lay-out: own address at the bottom of the letter, recipient, date at the top of the letter! Salutation: dear sir/madam, closing sentence, name and signature
Language: sentence construction good,
use of words generally very good! take note: "if the idea could proceed…" what do you mean exactly? Grammar, spelling, punctuation: good
Structure: show a clear introduction – core – conclusion with blank lines!
Conclusion: write good closing line! e.g., thank you in advance for sending the package of Smurfs.
Extra: rewrite! Take note: clear structure conclusion-core-intr and determine correct content! Good letter! Take note of the following: Everything left side, Salutation, Receipt from April 15th and campaign ends on April 20th!!
Proper use of language

and desired performance with regards to every aspect of writing. This enables a student to make more improvements, however, it can be questioned whether the student will use all feedback to improve the text and whether it transfers to other writing tasks as well. Further, teacher 2 responded with comments as well as with corrections in the text, whereas teacher 1 responded only with comments. For example, both teachers responded on the absence of the salutation and students' own name and address in the letter. Teacher 1 described this in a comment below the text, whereas teacher 2 added the necessary information already for the student on the right place in the text. The comments of teacher 2 were also highly directive: it rather explicated what information was lacking and how this should be included, than that suggestions or explanations were offered to help the student understand why this information should be included. This kind of directive feedback guides students on how to improve the text. However, especially for error corrections, it can be questioned whether it also improves the student as a writer as the suggested corrections can be adopted without any deeper understanding.

Figure 8.2b shows how the two teachers responded to a text of very high quality. It can be seen that both teachers provided feedback almost in the same way as they did for the text of very poor quality. In line with the quantitative findings, the examples show that both teachers responded somewhat more positive than they did in Figure 8.2a. If we take a closer look at their positive comments, it can be seen that teacher 1 indicated that the current performance is good (e.g., 'nice letter'), but that the feedback lacks specific information on why the letter is good. Teacher 2 explained more specifically what features of the writing are good (e.g., sentence construction is good, use of words is very good), but the positive text features become nearly invisible because of the large amount of feedback on features that still need improvement.

DISCUSSION

Feedback is generally regarded as an important instructional practice for learning to write, however, to reach its full potential, feedback should be adapted to the needs of the student. By tailoring feedback to students' individual needs, it can form the basis for enhanced learning for students with different proficiency levels (Wiliam, 2011). The present research investigated how upper-elementary teachers provide feedback on written products and whether they adapt their feedback to students' writing performance. We analyzed feedback on four dimensions, two related to the content of feedback (higher versus lower order feedback and positive versus negative feedback) and two related to the form (directive versus facilitative feedback and comments versus corrections). The results showed large differences between teachers' feedback practices, indicating that teachers had a preferred style of feedback. On average, teachers (1) provided an equal amount of higher and lower order feedback, (2) focused primarily on negative features in the text, (3) provided directions rather than facilitative feedback, and (4) communicated their feedback by textual corrections as well as by comments.

Further, teachers' feedback was hardly related to the quality of the written text: When teachers focused primarily on lower order concerns in the text or communicated their feedback with corrections in the text, they did so for poor as well as for high quality texts. The only influence of students' writing performance on teachers' feedback was found for the amount of praise and directive strategies, which was higher for texts of high quality than for texts of poor quality. From these results we can conclude that the content of feedback, as well as how it is provided, depends more on the style of the teacher than on students' writing performance.

The results from the present study suggest that teachers' feedback might not be effective in guiding students' writing development with one step at a time. Instead, teachers seem to be more focused on improving the text rather than improving the writer. First, the results showed that teachers were likely to respond to almost every error in the text, often without offering any explanation. Such detailed feedback does not increase students' knowledge about what constitutes good writing, as it does not focus students' attention to the main problems in the text (cf. Biber et al., 2011; Underwood & Tregidgo, 2010). Beginning writers are supported by feedback that prioritizes certain problems before others, as they often find it hard to evaluate the quality of their own text and to determine what needs to be improved (Chanquoy, 2001). This especially concerns the higher order aspects of writing (Sommers, 1980). Second, the feedback examples made clear that even when teachers provided positive feedback, it lacked explanation such that transfer to other tasks will be limited. Third, it was shown that teachers in the present study used more directive strategies for students who were already able to write a good text.

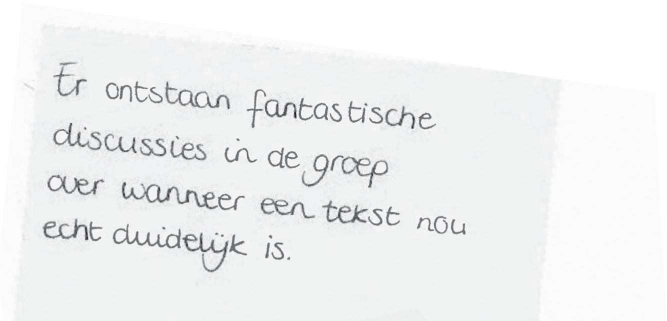
Taken together, these findings do not correspond with research on effective feedback emphasizing that teachers have to adjust their feedback to the students' individual needs (cf. Parr & Timperley, 2010). Hence, the results shed doubt on the effectiveness of teacher feedback for learning to write. However, whether feedback is effective can only be determined by examining its impact on students' writing performance. Follow-up research should therefore include revision as well as transfer tasks in order to understand what a student does with teachers' feedback.

The large differences between teachers in the present study suggest that teachers have a preferred style of providing feedback. In line with these findings, Ferris (2014) showed large variation in feedback practices of university teachers. She distinguished four general types of feedback givers: teachers who are concerned more with the writing process than with the end product, teachers who aim to provide feedback in the most efficient way, teachers who are very directive, and teachers who resist from taking over the ownership from students. The same typology seems to be applicable to the teachers in the present study. Moreover, the results in the present study have shown that teachers differ largely in the amount of feedback they provide, which can also be an important factor in distinguishing between different types of feedback givers. It would be interesting for future research to examine whether elementary teachers have a comparable style of providing feedback as teachers in secondary or higher education, and to examine how feedback styles relate to students' revision behavior.

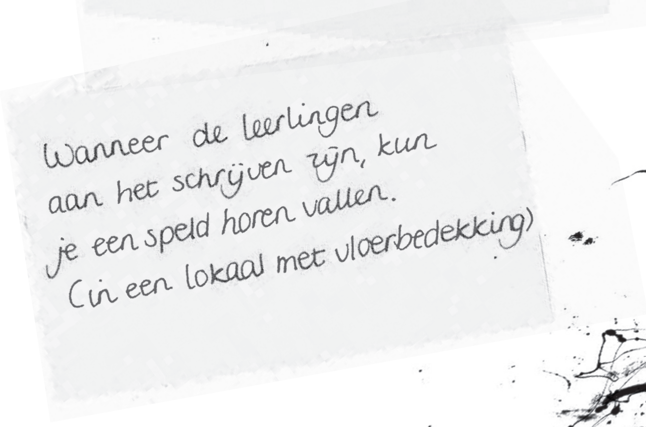
To improve the effectiveness of their feedback, teachers have to learn how to differentiate their feedback for weak and good writers. This is not an easy task for teachers, even for those who do believe that there is more to good writing than accuracy, and who believe that students should take more responsibility in text revision (Lee, 2008). First, teachers have to get an overall idea of the quality of a text; which is quite hard because text quality is not something which is simply right or wrong (Huot, 1990b). Second, teachers have to evaluate which elements are good in the text and what needs to be improved. Third, they have to select only the major points of feedback and decide how much support a particular student needs to revise the text on this point (Brannon & Knoblauch, 1982; Straub, 1996). In other words, teachers need both content and pedagogical knowledge to provide effective feedback (Parr & Timperley, 2010). An instrument that makes the learning trajectory for writing visible by providing examples of texts of poor and high quality might support teachers in deciding how far students' current performance is distanced from the desired performance. This can help them to focus only on main issues to improve students' writing with one step at a time (Parr & Timperley, 2010). Such a procedure to improve teachers' feedback practice should be developed and tested in future research.

In the present research teachers provided feedback to texts outside their actual classrooms. Fife and O'Neill (2001) argued that for understanding teachers' feedback practice, students' perspective should be taken into account. A similar point is made by Murphy (2000), who calls for more attention for how students react to and interpret teacher comments. However, the teachers in the present study responded in such a consistent way regardless the quality of the written text that it is rather unlikely that they would have responded in a total different way when the texts would have been from their own students. Further, knowing student's preferences or student's background might also impact the evaluation and feedback in a negative way. Teachers sometimes find it hard to provide feedback on a student's text, especially when they know that a student worked very hard to produce a text. To understand the interaction between teachers and students, follow-up studies should be conducted in teachers' regular classrooms.

To conclude, writing is a complex task for students in elementary school. Learning to write is even more challenging, as students have to write and, at the same time, have to learn from this to further improve their writing skills (Rijlaarsdam & Couzijn, 2000). Feedback has the potential of being a highly effective tool to teach students to write; not only because they have already written a text and can devote all cognitive capacity to learning from the feedback, but also because feedback provides the ultimate opportunity for teachers to differentiate their instruction to individual needs, providing the optimal possibility to enhance student's self-regulation. The current study shows, however, that teachers have their own style of feedback and that the content and form of their feedback is hardly adjusted to the quality of the written text. By relating text quality to key dimensions of teachers' feedback, this research offers important insight in how teachers can optimize their feedback for learning purposes.



Er ontstaan fantastische
discussies in de groep
over wanneer een tekst nou
echt duidelijk is.



Wanneer de leerlingen
aan het schrijven zijn, kun
je een speld horen vallen.
(in een lokaal met vloerbedekking)

Chapter 9

PROFESSIONAL DEVELOPMENT OF TEACHERS IN THE IMPLEMENTATION OF A STRATEGY-FOCUSED WRITING INTERVENTION PROGRAM FOR ELEMENTARY STUDENTS

This chapter is based on: Koster, M., Bouwer, R., & Van den Bergh, H. (2016). Professional development of teachers in the implementation of a strategy-focused writing intervention program for elementary students. Manuscript is revised and resubmitted for publication.

In this study we examined the effectiveness of Tekster, a comprehensive program for writing for the upper elementary grades, combining strategy instruction, text structure instruction, and the teaching of self-regulation skills with observational learning, explicit instruction, and (guided) practice to address both the focus of instruction (what is taught) and the mode of instruction (how it is taught). Further, we investigated the added value of a professional development program for teachers on the effectiveness and implementation of the intervention in the classroom, by adopting a teachers-training-teachers approach. One group of teachers (N=31) was trained by experts, and subsequently trained their colleagues (N=37). Quasi-experimental results showed that students' writing performance improved after the intervention (ES = 0.55), while generalizing over tasks, students, and teachers. Further, teachers became more positive and felt more efficacious about teaching writing after the intervention. There were no differences between trainers and trainees, which provides evidence for the spillover effect of professional development. To get more insight in how teachers implemented the intervention in their classroom and in the social validity of the intervention and the teachers-training-teachers approach, we triangulated post-intervention questionnaires with classroom observations and interviews. This mixed methods approach revealed that both trainers and trainees were highly satisfied with the program and easily adapted their focus of instruction. However, for adjusting the mode of instruction more teacher support seems to be needed.

INTRODUCTION

It is a major cause for concern that in the Netherlands, like in many other countries, students' writing performance at the end of elementary school does not meet the standards set by the Ministery of Education (cf. Department for Education, 2012; Henkens, 2010; Salahu-Din, Persky, & Miller, 2008). As a target goal for the end of elementary school the Ministry proposes that "students are able to write coherent texts, with a simple linear structure on various familiar topics; the text includes an introduction, body, and ending" (Expert Group Learning Trajectories, 2009, p.15). However, at the end of elementary school the majority of Dutch students is not capable of composing a text that successfully conveys a message to a reader (Kuhlemeier, Til, Hemker, De Klijn, & Feenstra, 2013). Why is writing so hard for elementary students? The major problem developing writers face during writing is cognitive overload. Writing is a complex cognitive process, during which several resource-demanding cognitive activities have to be performed simultaneously, such as activating prior knowledge, generating content, planning, formulating, and revising, whilst taking into account the communicative goal of the text and the intended audience (Fayol, 1999). Additionally, the amount of attention required for foundational skills (e.g., handwriting, spelling, and sentence construction) needs to be considered. This is particularly relevant with developing writers, as they often lack automaticity in these areas (McCutchen, 2011). Due to this limited automaticity, the learner has less attentional capacity for the higher level processes in writing, such as planning, formulating, and revising, which has detrimental effects on text quality (Berninger, Yates, Cartwright, Rutberg, Remy, & Abbott, 1992; McCutchen, 1996). An additional source of cognitive overload is the fact that, in the way writing education is often organized, learning-to-write and task execution are inextricably linked. For novice writers text production is already so cognitively demanding, that there is hardly any attentional capacity left for learning (Rijlaarsdam & Couzijn, 2000). Thus, writing instruction should aim to improve students' writing performance by teaching them skills and knowledge to manage the cognitive activities during writing. To achieve this, writing instruction needs to address the focus of instruction (what is taught) as well as the mode of instruction (how it is taught).

Writing instruction: The present situation

The Dutch Inspectorate for the Education reported that in the average classroom attention and time devoted to writing are limited, and that the majority of teachers do not succeed in effectively teaching writing (Henkens, 2010). There are two reasons for these shortcomings in writing education: (1) a lack of suitable teaching materials, and (2) teachers lack the necessary skills and knowledge for effectively teaching writing (Pullens, 2012; Van der Leeuw, 2006). Teachers often do not explain how students can approach a writing task, discuss texts, provide feedback, nor do they promote rereading and revising activities (Henkens, 2010). Although the language teaching materials pay attention to process-directed writing education, they do not offer teachers enough support to adequately assist their students during the writing process. Support for teachers is essential, as during their preservice and in-service professional development they are not sufficiently prepared to teach writing (Pullens, 2012; Van der Leeuw, 2006). Time devoted to the didactics of writing is limited, and student-teachers are expected to acquire the required skills and knowledge independently through learning-by-doing. As part of their training prospective teachers have to write a lot, but due to limited time and resources, they hardly receive any feedback on their writing (Van der Leeuw, Pauw, Smits, & Van de Ven, 2010). Thus, not only teaching materials need to be improved, but also the skills and knowledge of teachers need to be extended to optimize the focus and mode of writing instruction in elementary school. Already a lot of research has been done on both these aspects, identifying several effective instructional practices. These will be discussed below.

Optimizing the focus of instruction

Concerning the focus of instruction, several meta-analyses have identified various effective instructional practices to enhance students' writing performance, such as strategy instruction, teaching students self-regulation skills for writing, and text structure instruction, (Graham, 2006; Graham, McKeown, Kiuhara, & Harris, 2012; Koster, Tribushinina, De Jong, & Van den Bergh, 2015). Teaching students to adopt strategies before, during and after writing is an effective way to reduce cognitive overload during writing as this limits the number of cognitive processes that are active at the same time (Kellogg, 1988, 2008). For example, when students are taught to plan during the prewriting phase, they can focus on non-planning processes during writing. Studies involving explicit strategy instruction invariably yield large effect sizes, ranging from 0.82 to 1.15 (Graham, 2006; Graham et al., 2012; Graham & Perin, 2007; Hillocks, 1984; Koster et al., 2015).

The combination of strategy instruction with teaching self-regulatory skills yields an even higher effect size, ES = 1.17 (Graham et al., 2012). Essential self-regulatory skills in writing are setting goals for writing, and subsequently monitoring the progress towards these goals (Flower & Hayes, 1981). The most prominent and well-researched approach combining strategy instruction and the teaching of self-regulation skills is the Self-Regulated Strategy Development (SRSD) (Harris, Graham, Mason, & Saddler, 2002). In SRSD students are taught strategies for planning, writing, revising and editing, and they are supported in the development of the self-regulation procedures needed to monitor and manage their writing. This instructional approach has been implemented in small groups and whole classrooms with students of different age groups and abilities, and has invariably proven to be very effective in improving students' writing performance (Harris et al., 2002).

To be able to set effective goals for writing, students need to know what communicative goals should be set for which type of text and how you write a text meeting these goals. For this, students need to have knowledge about text structures and criteria for a good text. The effect of explicit text structure instruction, in which the elements and organization of text types are specifically taught, has been extensively examined in the elementary grades, in different genres: narrative (Fitzgerald & Teasley, 1986; Gordon & Braun, 1986), persuasive (Crowhurst, 1990, 1991; Scardamalia & Paris, 1985), and informative (Bean & Steenwyk, 1984; Raphael & Kirschner, 1985). Meta-analyses (Graham et al., 2012; Koster et al., 2015) show that the overall effect of text structure instruction was positive (ESs 0.59 and 0.76 respectively).

Optimizing the mode of instruction

Writing instruction must be optimized to address the double challenge problem of learning-to-write and task execution. An effective approach to separate these two components and provide students with the opportunity to fully direct their attention to learning-to-write is observational learning (Zimmerman & Risemberg, 1997). By observing a model performing (part of) a writing task while explaining, demonstrating, and verbalizing his thoughts, students gain insight into the writing process. This prepares them for the writing task and supports them during their writing process (Rijlaarsdam & Couzijn, 2000). Various studies have demonstrated the effectiveness of teacher modeling as an instructional mode to teaching writing strategies (cf. Graham, Harris, & Mason, 2005; Fidalgo, Torrance, Rijlaarsdam, Van den Bergh, & Lourdes Álvarez, 2015). Peers can also be used as models (cf. Braaksma, Rijlaarsdam, Van den Bergh, & Van Hout-Wolters, 2010). Besides positive effects on students' writing performances and writing processes (Braaksma, 2002; Braaksma et al., 2010), peer modeling also has positive effects on self-efficacy and motivation, especially in weaker students (Schunk, 1987).

Observational learning can also be applied by confronting students with reader reactions to provide them feedback on the communicative effectiveness of the written product (cf. Couzijn & Rijlaarsdam, 2004; Holliway & McCutchen, 2004). Beginning writers often are unaware of the communicative deficiencies in their writing. Observing genuine readers and discussing readers' experiences provide students with valuable information on the readers' needs and whether they succeeded in fulfilling these needs (Couzijn & Rijlaarsdam, 2004; Schriver, 1992). Several researchers (Couzijn, 1995; Couzijn & Rijlaarsdam, 2004; Holliway & McCutchen, 2004; Rijlaarsdam, Couzijn, Janssen, Braaksma, & Kieft, 2006) have shown that students' writing improved when they experience the effect their text has on a reader.

Although observational learning is effective in improving students' writing, there is still a gap to be bridged: from observing to independent practice. The teacher can facilitate student's progress through scaffolding with a gradual release of responsibility. In scaffolding the teacher controls the elements of the task that are initially beyond the student's capacity, thus permitting the student to concentrate upon the elements that are within his range of competence (Wood, Brunner, & Ross, 1976). The amount of teacher assistance can gradually be decreased as the learner progresses, and through guided practice and, finally, independent performance the cognitive load shifts from teacher to student

(Pearson & Gallagher, 1983; Wood et al., 1976). Intervention programs that use gradual release of responsibility and scaffolding have been successful in improving students' writing performance (cf. Graham, MacArthur, & Schwartz, 1995; Graham et al., 2005).

Bringing writing research into the classroom

Teacher involvement in research. Although the last decades of writing intervention research have provided guidance for the improvement of the teaching of writing, the actual implementation of evidence-based instructional practices is arduous, due to a substantial gap between research and classroom practice (Broekkamp & Van Hout-Wolters, 2007). To bridge this gap and effectively improve classroom practice, it is essential to involve teachers in intervention studies in a meaningful way (Borko, 2004). In the intervention studies ($N = 32$) that were analyzed in the meta-analysis of Koster and colleagues (2015), regular classroom teachers were not involved in the research in nearly half of the studies; the intervention was delivered either by the researchers themselves, or by trained research assistants. The results of these studies show positive effects on students' performance, but one can hardly expect any improvement after the intervention, as there is no encouragement or support for teachers to change the way they teach writing. This is also the case for 12% of the studies in which the teacher delivered the intervention with materials supplied by the researcher, without any additional training. Once the intervention has ended, teachers will return to working with their regular materials, so here also no longer lasting intervention effects are to be expected.

In the remaining 40% of the sample, teachers deliver the intervention themselves, either after having received a (short) training (21%) or a more extensive form of professional development (12%). Lastly, in 6% of the studies the teacher is part of the research team. Only under these circumstances, a change of teachers' instructional practices may be expected. It is worth mentioning that in this sample of studies the average effect size does not differ significantly between studies in which teachers were involved (ES = 0.79, $SD = 0.86$) versus studies in which they were not (ES = 0.99, $SD = 0.74$, $t(31) = .70$, $p = .49$). Thus, the inclusion of teachers in research does not seem to lead to a significant decrease in effect sizes. This pleads for inclusion of teachers in research.

Professional development. To meaningfully involve teachers in intervention research and improve classroom practice, teachers should be provided with the prerequisite tools to successfully implement the intervention. Therefore it is important that intervention studies include professional development activities for teachers. An effective way to organize professional learning for teachers is the practice-based professional development approach (Ball & Forzani, 2009). Practice-based professional development focuses on developing teachers' understanding and skills to effectively implement an educational practice, instead of focusing primarily on increasing teachers' knowledge about a practice (Ball & Forzani, 2009). Important features that have proven to be effective for teachers' professional development are (a) consistency with existing knowledge and beliefs, (b) focus on content and how students learn that content, (c) alignment with state standards, (d) opportunities for teachers to engage in active learning and (e) collaboration between teachers (Desimone, 2009; Harris, Lane, Graham, Driscoll, Sandmel, Brindle, & Schatschneider, 2012). Collaboration can be promoted by collective participation of teachers in schools, and by providing time and space for sharing, observing expert teachers in practice, being observed and receiving feedback, for instance in professional learning communities (Borko, 2004; Guskey, 1994; Harris et al., 2012). It has been established that professional learning communities lead to increased involvement, ownership, innovation, and leadership among teachers (Borko, 2004). Finally, professional development activities or programs should be sufficient in duration: this concerns the actual number of hours spent as well as the time span over which the trajectory is spread (Desimone, 2009).

Teachers' self efficacy for teaching writing. Improved teacher practices have a positive influence on student achievement (Desimone, 2009). As it is important that teachers are confident that they can affect students' learning outcomes, professional development should not only address the skills and knowledge that are required for successful and effective writing instruction, but also the beliefs about and attitudes toward writing instruction (Graham, Harris, Fink, & MacArthur, 2001). The beliefs that teachers hold about their ability to teach writing affect how they use the skills and knowledge that they have about teaching writing during their writing instruction (Graham et al., 2001; Pajares, 1992; Rietdijk, Van Weijen, Janssen, Van den Bergh, & Rijlaarsdam, 2015), which influences the overall quality of the instruction (Tschannen-Moran, Woolfolk Hoy, & Hoy, 1998). De Smedt, Van Keer, and Merchie (2016) found that teachers' efficacy for writing was positively related with students' writing performance. Thus, a higher feeling of self-efficacy of teachers for (teaching) writing results in a higher quality writing instruction, which leads to better student performance. This suggests that teachers' self-efficacy is a key factor in the improvement of writing education and that to improve the quality of teachers' instruction it is essential to enhance their feeling of self-efficacy by training them in applying effective writing practices (De Smedt et al., 2016).

Social validity. Another essential aspect in bridging the gap between research and practice is the social validity of an intervention, i.e., the acceptability of and satisfaction with the intervention procedures according to the individuals who receive and implement the intervention procedures (Luiselli & Reed, 2011). In the case of an intervention aimed to improve classroom practice, teachers might be queried about the complexity of the followed procedure, time involved with the implementation of the intervention, and satisfaction with the outcome (Luiselli & Reed, 2011). This yields important information about the feasibility of the implementation of an intervention in daily classroom practice. Procedures that are perceived as too complicated, impractical or unhelpful, will likely not be adopted. Social validity is a key aspect in the long term effects of interventions: a higher social validity increases the likelihood that (aspects of) an intervention will still be applied after the intervention period has ended. Social validity can be assessed through interviews, surveys, or questionnaires. However, each of these measures separately only provides information about a single aspect of the intervention. To obtain an impression of the full potential of the usability of an intervention in daily classroom practice, the results of the various social validity measures should be combined, for instance by using a mixed methods approach (Luiselli & Reed, 2011).

Testing the effectiveness of a writing intervention program in the classroom

The need to include professional development activities for teachers in the implementation of an intervention program was illustrated by the results of a recent intervention study, in which the effectiveness of a newly developed comprehensive program for teaching writing was examined (Bouwer, Koster, & Van den Bergh, 2016a). This program, called Tekster [Texter] (Bouwer et al., 2016a; Koster, Bouwer, & Van den Bergh, 2014a, 2014b, 2014c), aimed to improve the writing performance of students in the upper grades of elementary school in the Netherlands, and combined several effective instructional practices into one general overall approach for writing. In Tekster the main focus of instruction was to teach students a strategy for writing, based on the steps of the writing process, i.e., planning, writing, and revising. The main focus in grade 4 was on prewriting activities (generate and organize content), this shifted to post-writing activities (evaluating and revising) in grade 6. Strategy instruction was supplemented with explicit instruction in text structure and the teaching of self-regulatory skills. The predominant mode of instruction was observational learning, complemented with explicit instruction and (guided) practice with extensive scaffolding, following the gradual release of responsibility model.

Tekster was tested in a large intervention study in a natural setting, with 60 teachers and 1420 students. The intervention was delivered by the teachers themselves, after only one short introductory training session. Results showed that the program was already effective over a period of two months: students' writing performance improved significantly across all grades (ES = 0.40), whilst generalizing over tasks, and this improvement was maintained two months after the intervention (Bouwer et al., 2016). However, the intraclass correlation was .35, indicating that there was a large proportion of variance attributable to classes: in some classes students hardly made any progress in their writing performance, whereas in other classes students progressed a full grade level. These large differences between classes indicate that teachers need more support in implementing the program more effectively and with more fidelity. A supplementary professional development program could offer the support teachers need and might minimize differences between teachers.

However, in a large-scale intervention study it is not always a feasible option to have all teachers involved in the study participate in an extensive professional development program. It is therefore necessary to examine alternative ways in which teachers can benefit from professional development. Sun, Penuel, Frank, Gallagher, and Youngs (2013) have investigated the so-called spillover effect, which means that through collegial interactions teachers can learn from professional development participants. The results from this study are promising: spillover effects can be almost as large as the direct effects of professional development. In the US National Writing Project (NWP) a similar method was used to reach large numbers of teachers: a teachers-training-teachers approach (Borko, 2004; Lieberman & Friedrich, 2007). In this project, teachers attended summer institutes and subsequently provided workshops for their colleagues. These teacher-trainers reported that it took time to earn their leadership, but that they eventually succeeded in doing so by showing their commitment, by high-quality teaching, and their willingness to give advice to their colleagues (Lieberman & Friedrich, 2007). Teachers reported that NWP has brought about change in their beliefs and attitudes about teaching writing, and the majority of students' work showed improvements in organization, coherence and use of writing conventions (Borko, 2004). Thus, a teacher-training-teacher approach is a promising method to train large numbers of teachers.

Aim of the present study

In the present study, we examined the effect of the writing intervention program Tekster, including professional development activities for teachers, on students' writing performance, and on teachers' self-efficacy and attitudes for writing and the teaching of writing. Results from a previous intervention study showed that the program was effective, but large differences between teachers were an indication that more support in implementing the program was required (Bouwer et al., 2016a). Therefore, in this study, we included professional development activities to offer teachers support in the implementation of Tekster and increase their skills and knowledge of teaching writing. To investigate whether the content of such a professional development program could be transferred between teachers, we applied a teachers-training-teachers approach in which half of the participating teachers were trained by the researchers, and these teacher-trainers subsequently trained one or more colleagues.

We investigated the effect of the intervention on student outcomes, and examined whether there were differences between students in the group of teachers who were trained by experts (teacher-trainers), and teachers who were trained by colleagues (teacher-trainees), and whether there were differences between grades. Further, we examined the effect of the intervention program and professional development activities on the self-efficacy and attitudes of teachers towards writing and teaching writing and whether there were differences between teacher-trainers, and teacher-trainees. The effect of the intervention on students' writing performance was examined by a quasi-experimental design with two groups (teacher-trainers versus teacher-trainees) and three measurement occasions. The first measurement occasion was a pretest for both groups. The first group (teacher-trainers) received the intervention during the first time interval, from the first to the second measurement occasion. The second measurement occasion served as a posttest for this group, and as a second pretest for the second group (teacher-trainees). Between the second and third measurement occasion, the second group received the intervention. Therefore, the third measurement occasion served as a posttest for the second group, as well as a delayed posttest for the first group, which provided information on the long-term effects of the intervention. To examine the effect of the intervention on the self-efficacy and attitudes of teachers, and whether this effect differed between teacher-trainers and teacher-trainees, we administered questionnaires prior to and after the intervention.

We expected that students' writing would improve after the intervention in both groups, but not necessarily to the same extent, as research has shown that spillover effects can be almost as large as the direct effects of professional development (Sun et al., 2013). Regarding the differences between grades we expect that students in grade 6 will write qualitatively better texts than students in grade 4, as they have had more years of schooling and practice. We expect improvements in self-efficacy, and that all teachers become more positive about writing and writing instruction due to the professional development program, but again, not necessarily to the same extent.

As teachers were a crucial factor in the implementation of the intervention program, it was important to establish how they implemented the program and if they implemented the program with fidelity. For this, we specifically examined how teachers implemented key components of the intervention program in their writing lessons and how trainers transferred their know-ledge and skills to colleagues during the collegial training sessions. Lastly, we investigated the social validity of the intervention program and the teachers-training-teachers approach, as this provides valuable information about the feasibility of the implementation of the intervention in daily classroom practice. With regards to social validity, we examined teachers' experiences and satisfaction with the lesson material, the teacher manual, and the training sessions. To examine the implementation and the social validity of the intervention program including the professional development activities, we used an explanatory sequential mixed methods approach in which we collected and analyzed quantitative data from post-intervention questionnaires, logbooks and classroom observations, and followed up with qualitative data from focus group interviews to further explain how teachers implemented and experienced the intervention program and to elaborate on differences between trainers and trainees. This will provide us with valuable clues on how the program was implemented in schools.

METHOD

Sample

In total, 68 teachers and 1365 students from 65 classes and 25 elementary schools participated. All teachers were qualified and experienced elementary teachers. On average, they had 11.22 years ($SD = 8.50$) of experience in elementary grades, with an average of 3.55 years in the grade in which they were teaching at the start of the study. The majority of teachers were female (81%) and they were from schools spread all over the country: 10 schools were located in the northern region, 9 in the middle region, and 6 in the southern region. Schools varied in their identity: 14 schools were grounded in a religious denomination and 9 schools applied innovative teaching concepts, such as Montessori or Dalton. There were 20 fourth grade classes, 14 fifth grade classes, 20 sixth grade classes, and 11 multigrade classes combining two or three grades participating in the experiment.

Teachers, volunteering to participate in the study ($N = 30$), were assigned to the teacher-trainer group. This group was provided a professional development program in writing education by the researchers. These teachers had to bring at least one colleague who they trained themselves for the duration of the study. So, the sample included at least one teacher-trainer and one teacher-trainee from each school. In total, there were 38 teachers in the teacher-trainee group. Teacher-trainers had slightly more teaching experience than teacher-trainees ($M = 13.83$ years, $SD = 8.20$ versus $M = 8.98$ years, $SD = 8.21$, $t(63) = 2.38$, $p < .05$), but they did not differ significantly in years of experience in the grade in which they were currently teaching ($M = 3.51$, $SD = 3.59$, $t(63) = 1.51$, $p = .14$).

Table 9.1 presents the number of students per grade in each group. The trainer group consisted of 28 classes with a total of 602 students. The trainee group consisted of 37 classes with a total of 763 students. As can be seen in Table 9.1, there were no differences between the two groups in the percentage of female and male students ($\chi^2(2) = .25$, $p = .62$). Students' age ranged from 8 to 13 years, with an average of 10.23 years ($SD = 1.00$), which did not differ between groups ($F(2) = 0.81$, $p = .37$). There were minor differences in the language background of the students of both groups ($\chi^2(2) = 16.73$, $p < .001$). Dutch was the native language for most students in the trainer (64%) and trainee group (74%), but the trainer group consisted of relatively more students for which Dutch was the L2 (36%) than the trainee group (26%). The most frequently spoken languages besides Dutch were Arabic, Turkish, English and Frisian (a language spoken in the northern regions of the Netherlands).

In total there were 19 students who dropped out during the study because they changed schools: 11 students in the trainer group and 8 students in the trainee group. These students were removed from the data set, which resulted in a total sample of 1346 students.

Table 9.1 Student information per grade and training group

	Trainer group			Trainee group		
Grade	*N* students	% female	Mean age (*SD*)	*N* students	% female	Mean age (*SD*)
4	178	49	9.19 (0.49)	283	49	9.16 (0.49)
5	197	49	10.28 (0.51)	230	50	10.24 (0.55)
6	227	49	11.24 (0.56)	250	50	11.23 (0.52)
Total	602	49	10.32 (0.99)	763	50	10.16 (1.01)

Design

The intervention in this study is implemented by using a design with switching panels (Shadish, Cook, & Campbell, 2002), with two groups (trainers and trainees) and three measurement occasions. In this design the intervention is implemented in both groups, but at different moments in time. An advantage of this design is that, as the intervention is implemented consecutively in the two groups, it is possible to test whether the effectiveness of the intervention differs between teacher-trainers and teacher-trainees. If the effect of the intervention is equal for both groups, this indicates that professional development provided by experts or colleagues is equally effective, meaning that professional development can be transferred between colleagues and does not rely on experts. Another advantage of a switching panel design is that all students eventually benefit from the intervention, making it a more ethical design than a regular pre-post (quasi-) experimental design, see also Bouwer et al. (2016a).

Teachers and students in the trainer group started with the intervention program in the first period, between the first and second measurement occasion. This period lasted four months during which teachers executed one writing lesson a week instead of their regular program for writing. While the trainer group started with the intervention, the trainee group served as a control group, engaging in their regular writing activities and routines. Whereas the second measurement occasion for the trainer group was scheduled after four months (after completing the whole intervention program), it was scheduled already after two months for the trainee group. This overlap in time was created for an optimal implementation of the teachers-training-teachers approach, as jointly working on the program would promote collaboration and interaction between trainers and trainees (Borko, 2004).

After the second measurement occasion, teachers and students in the trainee group started with the intervention, while teachers and students in the trainer group returned to their regular writing activities. The procedure for the trainee group was the same, teachers executed one Tekster-lesson a week for a period of four months. The only difference between the two groups was the professional development of teachers. Teachers in the trainer group received two expert training sessions (before and during the intervention program), whereas teachers in the trainee group were trained by their expert-trained colleagues. The third measurement occasion served as a posttest for students in the trainee group, as well as a delayed posttest for students in the trainer group, with which we were able to measure retention.

Regular writing education

In the present study, the intervention program is compared to the regular classroom practice in writing education. In the Netherlands, writing education is traditionally part of the language teaching curriculum. A report from the Dutch Inspectorate of Education (Henkens, 2010) showed that from the 8 hours per week reserved for language teaching, on average 45 minutes are spent on writing. These writing lessons are primarily product-focused: students hardly receive any support during the writing process, nor are they supported on how to approach writing tasks. In the majority of schools the writing performance of students is not monitored, and students are seldom given feedback on their performance. Questionnaire data on the regular classroom practice of the teachers participating in the present study show comparable results. A majority of teachers (70%) indicated to use textbooks for language teaching for their writing lessons. On average, they devoted 42.42 minutes ($SD = 29.91$) a week to writing in class, of which 13.93 minutes ($SD = 12.11$) were devoted to instruction in writing strategies. Teachers provided their instruction mostly plenary, less time was devoted to small-group instruction or individualized instruction. Overall, there were no differences between trainers and trainees in their regular writing practice ($p > .06$).

Writing intervention: Tekster

The intervention program consisted of a teaching program, Tekster, which included three lesson series of 16 lessons, one for each grade level, compiled in a workbook for students, accompanied by a teacher manual (Koster et al., 2014a, 2014b, 2015c). Further, to foster the professional development of teachers and support them in the implementation of the program two additional training sessions were provided. In Tekster the main focus of instruction was to teach students a strategy for writing, supplemented with explicit instruction in text structure and the teaching of self-regulatory skills. To support students in applying the writing strategy, they were taught a mnemonic representing the steps of the writing process: VOS (fox) for grade 4, DODO (dodo) for grade 5, and EKSTER (magpie) for grade 6. The letters of the acronyms represent the steps in the writing process as follows: VOS (fox) for Verzinnen (generate content), Ordenen (organize), Schrijven (write); DODO (dodo) for Denken (think), Ordenen (organize), Doen (do), Overlezen (read); Ekster (magpie) for Eerst nadenken (think first), Kiezen & ordenen (choose & organize), Schrijven (write), Teruglezen (reread), Evalueren (evaluate), Reviseren (revise). In the first lesson of the program the acronym-animal was introduced in a story in which students also practice the steps of the strategy for the first time. In the following lessons the animals are the common theme, with small icons of the animals serving as a visual support.

Table 9.2 gives a general description of the design principles of the program and how these principles were operationalized in learning and teaching activities (see Rijlaarsdam, Janssen, Rietdijk, & Van Weijen, 2015). In particular it shows how effective practices for mode and focus of instruction are combined and translated into learning activities for students and teaching activities for teachers, as well as the support that is provided for teachers in the teacher manual and during the training sessions of the professional development program.

General lesson format. To ensure that all activities described in Table 9.2 were covered, all Tekster-lessons were designed using a general (more or less fixed) lesson format. The core of the lessons was the overall writing strategy. Each lesson started with a plenary introduction in which the goal of the lesson is explicitly stated. Specific characteristics of the text type were addressed through modeling (teacher modeling, or peer modeling using videoclips), comparing model texts, or explicit instruction. Next, the authentic writing assignment was introduced, with explanation of the communicative goal and intended audience, and the acronym for the strategy was explicitly named. Subsequently, students started with the first step of the strategy, which was generating content in keywords, followed by the second step, which was organizing the generated content, supported by the teacher through scaffolding with gradual release of responsibility. In the third step of the lesson students started writing their texts using the organized content, while the teacher provided support when necessary. In the following step (grade 5 and 6 only) students read each other's texts or their own text. In the fifth step of the lesson (grade 6 only) students evaluated the written text by answering evaluative questions and/or giving feedback. In the sixth step of the lesson (grade 6 only), students revised (parts of) their text on the basis of the feedback they received. The duration of the average Tekster-lesson was between 45 and 60 minutes. A sample lesson is included in Appendix M.

Writing tasks in the lesson program. In the program students learn to apply the writing strategy to various types of texts, for which authentic writing tasks with various communicative goals and audiences are used. For instance, in each grade they learn to write descriptive texts (e.g., a self-portrait or personal ad), narrative texts (e.g., a story or newspaper article), persuasive texts (e.g., a nomination email for a television program or a flyer for recruiting new members for a club), instructive texts (e.g., a recipe, rules for a game) and personal communication (e.g., a holiday postcard or invitation). The writing tasks comply with the goals set for the end of elementary school by the Ministry of Education.

The level of difficulty ascended through the grades as follows: in grade 4, predominantly writing tasks were used in which the intended audience was in close proximity of the student, such as classmates, friends, or (grand-) parents. In grade 5, this was expanded to people with whom students have a more distal relationship, but are still familiar to them, such as their teacher, relatives, or neighbours. In grade 6, students also have to write texts that are intended for unfamiliar people, such as the editor of a newspaper, or the managing director

Table 9.2 Design principles for focus and mode of instruction of the teaching program Tekster and their translation into learning and teaching activities

Design principles		Teaching program		Professional development
Focus of instruction	**Mode of instruction**	**Learning activities**	**Teaching activities**	**Training & support for teachers**
1. Process related: Writing strategies *General approach for writing tasks, based on phases of the writing process: generate content, organize, formulate, reread, evaluate, revise*	a. Observational learning	Observe/discuss/compare model(s), (teacher or peer) applying the writing strategy in different stages of the writing process	Modeling strategy use (thinking aloud while performing (part of) the writing task)	DVD with example videos for modeling in the teacher manual, practicing modeling during first training session, instructions for the use of modeling in lesson plans in teacher manual
	b. Explicit instruction	Listen actively, retrieve relevant background knowledge from memory, take notes	Explain the components of the strategy, make students aware of the purpose and benefits of using writing strategy, activate student's background knowledge	Explanation of strategy and components in first training session and in general introduction of teacher manual, specific instructions for each lesson in lesson plans
	c. (Guided) practice	Apply the steps of the strategy to writing tasks: authentic tasks with clear communicative goal and intended audience in various genres	Provide help when needed through scaffolding and process feedback	Explanation of importance of feedback and practice with giving feedback in second training session, information on feedback in general introduction teacher manual, specific instruction for each lesson in lesson plans

Design principles		Teaching program		Professional development
Focus of instruction	**Mode of instruction**	**Learning activities**	**Teaching activities**	**Training & support for teachers**
2. Product related: Text structure *Criteria for written product, depending on communicative goal and intended audience*	a. Observational learning	Before writing: Observe/discuss/compare model(s), (teacher or peer) talking about criteria for and conventions of various text types, compare and discuss model texts of the same text type to derive criteria and conventions for a good text	Before writing: Model the relevant aspects of the text type, provide model texts or show video clips of peer modeling	DVD with example videos for modeling in the teacher manual, practicing modeling during first training session, instructions for the use of modeling in lesson plans in teacher manual, model texts are provided in workbooks, specific instruction for each lesson in lesson plans
		After writing: Evaluate peer/own text on the basis of the previously discussed criteria and give feedback (reader reaction), observe reader reaction, observe model revising on the basis of feedback	After writing: Evaluate students' texts on the basis of previously discussed criteria, give feedback (reader reaction), model how feedback can be used in revision to improve the text	DVD with example videos for modeling in the teacher manual, practicing modeling during first training session, instructions for the use of modeling in lesson plans in teacher manual, Explanation of importance of feedback and practice with giving feedback in second training session, information on feedback in general introduction teacher manual, specific instruction for each lesson in lesson plans

▼

Design principles		Teaching program		Professional development
Focus of instruction	**Mode of instruction**	**Learning activities**	**Teaching activities**	**Training & support for teachers**
2. Product related: Text structure *Criteria for written product, depending on communicative goal and intended audience*	b. Explicit instruction	Listen actively, take notes Apply the discussed criteria to writing tasks: authentic tasks with clear communicative goal and intended audience	Explain why and how the criteria and conventions should be used, discuss important criteria and conventions on the basis of model texts	Specific instruction for each lesson in lesson plans in teacher manual
	c. (Guided) practice	After writing: Give feedback /assess own text according previously discussed criteria	Provide help when needed through scaffolding and product feedback	Explanation of and practice with how to assess text quality of and provide feedback on students' texts in second training session, information on feedback in general introduction teacher manual, specific instructions and suggestions for each lesson in lesson plans, rating scales with benchmark texts for three genres (narrative/descriptive/argumentative)

Design principles		Teaching program		Professional development
Focus of instruction	**Mode of instruction**	**Learning activities**	**Teaching activities**	**Training & support for teachers**
3. Writer related: Self-regulation skills *Writer's monitoring and regulating of own progress in relation to communicative goals*	a. Observational learning	Observe/discuss/compare model(s), (teacher or peer) setting goals and monitoring progress in relation to goals during the writing process, observe/discuss/compare effect of self-regulation on the written product.	Model self-regulation during writing, by thinking aloud during performing writing task	DVD with example videos for modeling in the teacher manual, practicing modeling during first training session, suggestions for the use of modeling in teacher manual
	b. Explicit instruction	Listen actively, take notes	Explain why it is important to set communicative goals for writing in advance, explain the differences between various communicative goals, when and how during the writing process progress towards the communicative goal can best be monitored	General information on goal-setting and communicative goals in first training session and in introduction teacher manual. Specific information (i.e., communicative goal of the lesson) in lesson plans
	c. (Guided) practice	Set communicative goal before writing, monitor progress towards this goal during writing, regulate own writing process and adapt if necessary, evaluate written product in relation to communicative goal, revise if necessary.	Provide help when needed through scaffolding, and self-regulation feedback	Explanation of and practice with how to provide feedback in second training session, information on feedback in general introduction teacher manual, specific instructions and suggestions for each lesson in lesson plans

of a company. The writing tasks were developed in close collaboration with elementary teachers to ensure that the topics would match students' interest and developmental level. The teachers piloted the writing tasks in their own classrooms first, and subsequently, the program was tested in a pilot study (Koster, Bouwer, & Van den Bergh, 2016a) and in a large-scale intervention study (Bouwer et al., 2016a).

Professional development. The professional development component of the program consisted of a teacher manual and two training sessions.

Teacher manual. The teacher manual was provided to all teachers in order to facilitate the teaching of Tekster-lessons. The manual consisted of two parts: a general introduction and detailed lesson plans for each lesson. In the general introduction the goal and approach of the program were explicated as well as the general lesson format. Further, the importance of feedback for learning to write was explained and suggestions were given on how to provide effective feedback. Next, observational learning and modeling were explained, and practical information was given concerning the organization of the intervention. The detailed lesson plans provided an overview of the instruction and activities of the lesson with a time planning for each phase of the lesson. These lesson plans described the activities the teacher was expected to execute during the lesson and provided suggestions when to use modeling during instruction. The manual also included a dvd with movie clips of peer modeling to use during instruction, and videos with examples of teacher modeling for different phases of the lesson. Further, the manual included a benchmark rating scale to support teachers in evaluating text quality and giving feedback. The benchmark rating scale consisted of five students' texts of ascending quality, representative of the range of text quality that can be found in grade 4 to 6 (Bouwer, Koster, & Van den Bergh, 2016c).

Training sessions. To support teachers in the implementation of the program in the classroom, two training sessions were planned over the course of the intervention. During these sessions, which lasted four hours each, teachers were trained in small groups (max. 12 teachers) by the researchers. The righthand column in Table 9.2 demonstrates how the aspects of the training are related to the learning activities and the teaching activities of the program. During the first training session prior to the start of the intervention teachers were briefed on how to work with the program. The researchers informed the teachers about the theoretical background, and showed and discussed an example video of teacher modeling. Teachers were instructed to apply coping modeling during their lessons. In contrast to mastery models, who show a flawless performance, coping models initially display exemplary deficiencies but overcome these difficulties and gradually improve their performance (Schunk, 1987; Zimmerman & Kitsantas, 2002). Research has shown that observing coping models raised students' self-efficacy and enhanced their performance more effectively than a mastery model (Zimmerman and Kitsantas, 2002). This may be due to the explicit modeling of strategies to overcome difficulties, or it might be that, due to perceived similarity to the model, students believe that they are also able to improve their performance (Schunk, 1987). During the training session was discussed how teachers could implement modeling in their own writing lessons. Next, the teachers, in small groups, jointly prepared the first two lessons with special attention to where and how modeling could be applied in these lessons. Lastly, teachers were instructed to read the information in the manual carefully and watch the videos before the start of the program as a preparation for teaching the lessons.

The second training session was scheduled after six lessons. During this session, first experiences and specific issues regarding the implementation of Tekster in the classroom were shared and discussed. Next, teachers were trained in how to provide their students with effective feedback, and how to assess students' texts. For effective feedback teachers have to adjust their comments to students' needs, which requires that teachers are able to assess the quality of students' texts and adapt their feedback accordingly (Bouwer, Koster, & Van den Bergh, 2016b). Therefore teachers were trained in how to evaluate text quality using benchmark texts representing ascending levels of writing quality (Bouwer et al., 2016c). After this, they were introduced to the characteristics of effective feedback, followed by a plenary discussion of examples of teacher feedback. Subsequently they practiced how to provide effective feedback using the scale with benchmark texts: with example texts, but also with texts of their own students. They reflected on the quality of their feedback in subgroups.

Teachers-training-teachers approach. For the training sessions we adopted a teachers-training-teachers approach: teachers in the first group were trained by the researchers, subsequently these teachers trained their colleagues. Teachers who followed the professional development program received instruction and materials to subsequently train their colleagues, who started with the program two months later. The teacher-trainers were instructed to plan two training sessions with their colleagues, in which the same topics should be addressed as in the training sessions that they received. Further, it was encouraged that teacher-trainers would invite their colleagues in their classroom to observe a Tekster-lesson.

Measures and procedure

Assessment of students' writing quality. To examine how the intervention program affected students' writing performance, students completed three writing tasks prompting for different genres: descriptives, narratives and persuasive letters at each measurement occasion. The tasks within a genre were similar with regards to the communicative goal and intended audience, and only differed by topic. Hence, students wrote nine texts in different genres and topics, which warrants generalization to writing proficiency (Bouwer, Béguin, Sanders, & Van den Bergh, 2015). Each task contained a writing prompt, including an illustration with relevance to the topic, and some space for prewriting which students were free to use. Appendix G provides examples of writing prompts for each genre.

Similar writing tasks were used and validated in a previous study with students in the same age group (Bouwer et al., 2016a). Teachers administered the writing tasks to their students during normal class time, without providing any additional instruction. Students had to work individually on the task, without a time limit. Teachers were instructed to plan the three writing tasks for each measurement occasion within one week, but not on the same day.

The quality of students' texts was rated by eighteen experienced elementary teachers using a benchmark rating procedure. In this procedure, raters independently score each text by comparing it to a scale with five benchmark texts (Bouwer et al., 2016c). These benchmarks reflect the range of writing quality of students in grade 4 to 6. There were different benchmark scales for each writing genre, see Appendix K for an example. The center position on each scale is an average text which is assigned an arbitrary score of 100 points. The other texts on the scale are one (115 points) and two (130 points) standard deviations above average, and one (85 points) and two (70 points) standard deviations below average. This rating procedure was developed in a previous study in which its support for raters in assessing text quality across tasks and genres was demonstrated (Bouwer et al., 2016a; Bouwer et al., 2016c). To ensure that raters were blind to conditions, we anonymized students' texts. Each text was rated by a jury of three raters using a design of overlapping rater teams (Van den Bergh & Eiting, 1989). In this design, texts are randomly divided into subsamples, equaling the number of raters (N= 18). Subsequently, each rater received 3 subsamples according to a prefixed design. Because each subsample was rated by overlapping rater teams, it was possible to estimate the reliability of the scores of individual raters, and to approximate the reliability of jury raters (Van den Bergh & Eiting, 1989). The average reliability of jury ratings was high in the present study, overall $\rho = .88$, varying from $\rho = .83$ to $\rho = .90$ per task. The final text quality score was determined by averaging the scores of the jury raters. As scores appeared to be somewhat negatively skewed, raters' scores were normalized for each task using Blom's rank-based normalization formula (Solomon & Sawilowsky, 2009).

Teachers' self-efficacy and attitudes for writing. To gain insight into the influence of the intervention program and the professional development program on teachers' self-efficacy, attitudes towards writing and the teaching of writing, teachers filled in questionnaires prior and after the intervention program. Teacher efficacy for writing was measured by the Efficacy Scale for Writing (TES-W; Brindle, 2013; Graham et al., 2001). This scale measured teachers' beliefs about their own writing instruction on two dimensions: (1) the degree to which teachers attribute students' successful writing to their own writing instruction (3 items, e.g., "When students' writing performance improves, it is usually because I found better ways of teaching them") and (2) the perception of their ability to support inexperienced writers (4 items, e.g., "When I try really hard, I can help students with the most difficult writing problems"). The items were measured on a 5-point Likert scale ranging from 'strongly disagree' to 'strongly agree'. The scale has been validated in previous research (Brindle, 2013; De Smedt et al., 2016) and the internal consistencies for the subscales in the present study were satisfactory, respectively $\alpha = .77$ and $\alpha = .58$. Positive medium correlations between the subscales on pretest measures confirmed that the scales are related but measure different dimensions of teacher efficacy ($r = .39$, $p < .01$).

Teachers' attitudes for writing were measured by the questionnaire of Brindle (2013). The questionnaire included 4 items on teachers' attitudes towards writing (e.g., "I enjoy writing") and 4 items on their attitudes towards writing instruction (e.g., "I like to teach writing"), which were measured on a 5-point Likert scale ranging from 'strongly disagree' to 'strongly agree'. The scale has been validated in previous research (Brindle, 2013; De Smedt et al., 2016) and the internal consistencies for the subscales in the present study were high, respectively $\alpha =$ .84 and $\alpha = .89$. Positive medium correlations between the subscales on pretest measures confirmed that the scales are related but measure different dimensions of teacher attitudes ($r = .26$, $p < .01$).

Implementation and social validity of the intervention program. We used an explanatory sequential mixed methods design (Creswell & Plano Clark, 2011; Johnson, Onwuegbuzie, & Turner, 2007) to (a) examine whether trainers and trainees implemented the lesson program as intended, (b) investigate how trainers implemented the professional development program, and (c) get in-depth information about the social validity of the intervention. In the first phase we collected and analyzed quantitative data from logbooks, post-intervention questionnaires, and observations. The triangulation of the data of different measures provides a better understanding than either approach alone. Whereas questionnaires and logbooks provide general information on teachers' classroom practice and experiences, observations provide richer and more objective data about how teachers actually implemented the program (Desimone, 2009). In the second phase we collected qualitative data from focus group interviews. By exploring differences between trainers and trainees in more depth, the qualitative data further refine and explain the quantitative results obtained in the first phase (Creswell & Plano Clark, 2011). The specific qualitative and quantitative measures and procedures are further explained below.

Logbooks. Logbooks were incorporated in the teacher manual. Teachers were requested to fill in the logbook after each lesson. They were asked to provide the following information about each lesson in the program: preparation time, lesson duration, appreciation of the lesson (on a scale from 1 to 10; 1=very low, 10=very high), estimated level of difficulty for their students (on a 5-point scale; 1=easy, 5=hard), the level of difficulty of executing the lessons (on a 5-point scale; 1=easy, 5=hard), and additional comments (if any). After the intervention the logbooks were collected, together with the students' workbooks, to check whether all writing lessons were executed as planned.

Post-intervention questionnaire: Implementation of the lessons. To measure teachers' implementation of the lesson program Tekster, teachers were first asked to indicate whether they adapted lessons to their own context on an ordinal scale with three categories (never, sometimes, always), and to provide reasons for this. The next nine items measured how teachers implemented the key components modeling, feedback, and evaluating text quality. Three items measured how often during the intervention teachers modeled (a part of) the writing process, how often they provided feedback to students, and how often they evaluated students' writing products. This was measured on an ordinal scale with five categories: never, only once, in some lessons, every lesson, multiple times per lesson. Three items measured whether teachers used each component more frequently during the intervention than during their regular classroom practice for writing on a scale from 1 (less frequently) to 3 (more frequently). For the component feedback, they also had to indicate whether their feedback focused mainly on the writing product, the process, or both, and whether they provided mainly oral or written feedback. Further, they were asked to indicate whether they used a benchmark rating scale for evaluating text quality, providing feedback, and/or instruction in class.

Post-intervention questionnaire: Implementation of the training. To get more insight in whether the collegial training sessions were comparable to the expert training sessions, the questionnaire included five items measuring teachers' implementation of the teachers-training-teachers approach. Teachers were asked how often they organised a training session with their colleague(s), which was measured on an ordinal scale with four categories (never, once, twice, or more than twice), and they had to indicate the total training time in minutes. Further, they were asked whether the training was one-on-one or in a team setting. To get more information on the content of the collegial training sessions, we asked them to indicate whether they discussed the following topics: the goal and structure of the program, organisation of the lessons, modeling, feedback, rating text quality, the benchmark rating scale, student texts or specific issues. They also had to indicate whether they observed a Tekster-lesson of a colleague.

Post-intervention questionnaire: Social validity of the intervention program. The questionnaire further measured the social validity of the intervention program, including the lesson program and professional development activities. First, to measure teachers' satisfaction with the lesson program, trainers and trainees had to indicate their general attitude towards the lesson program Tekster on a 5-point Likert scale from 'highly negative' to 'highly positive'. They were also asked to indicate on a 5-point Likert scale from 'strongly disagree' to 'strongly agree' whether they believed that their students were supported by key components of the intervention, i.e., writing strategy, teacher modeling, peer modeling clips, model texts, feedback, and peer interaction. Further, they had to indicate whether they intended (or not) to keep using Tekster, components or specific lessons, after the study. Finally, to gain insight into students' satisfaction with the lesson program, we asked students to rate the overall program on a 10-point scale from 'very bad' to 'very good'. Teachers' experiences with the professional development activities were measured with six questions. Teachers had to indicate on a 5-point Likert scale whether their classroom practice was supported by the teacher manual, and by the benchmark rating scale that was included in the manual. Further, teachers had to indicate on a 5-point Likert scale whether the training sessions provided general support for their writing instruction, and specifically for modeling, providing feedback, and evaluating the quality of their students' texts. Together these questions provided information on the usability of Tekster in daily classroom practice and whether the knowledge and skills required for using Tekster effectively were easily transferable between colleagues.

Observations. Classroom observations generally provide more insight in the actual implementation of an intervention (Desimone, 2009). To obtain reliable and valid information on the actual implementation of the program in the classroom, observations of multiple lessons over an extended period of time are required (Desimone, 2009). Therefore we observed three lessons in the classrooms of one-third of the participating schools, which involved 22 teachers from 17 classes from 5 schools: 10 teacher-trainers and 12 trainees. We selected schools that were different in their identity and background in order to get a representative view on how Tekster is implemented in different contexts. Our selection included a large urban school with a large population of multicultural students for whom Dutch was a second language, a small rural school with a religious denomination, an urban school with a predominantly Dutch population, a school with multiple-grade classrooms, and a school where teachers work part-time. During the intervention period we observed multiple lessons in each of the 17 classrooms.

We observed lessons of trainers as well as trainees to investigate whether classroom practice differed between these groups. In total, we observed 24 lessons of trainers and 30 lessons of trainees. We used an observation instrument developed for a previous study in which the effectiveness of Tekster was examined (Bouwer et al., 2016a). The instrument consisted of two parts for each phase of the lesson. In the first part was tallied every 20 seconds whether the teacher was on task or off task: on task if the teacher was executing the actions as specified in the lesson plan for that particular phase of the lesson, off task if the teacher was involved in other activities than teaching writing, such as fetching a cup of coffee or talking to a colleague. Further, it was tallied whether the on-task-behavior involved plenary activities (instruction or classroom interaction) or interaction with individual students. Second, for every lesson phase observers had to register whether the teacher applied key components of the intervention, i.e., teacher modeling, referring to the acronym and/or the steps of the strategy, or providing feedback. Additionally, the time for each phase of the lesson and for the lesson as a whole was registered. Each classroom was observed by one trained undergraduate student, there were ten observers in total. To optimize observers' agreement, all observers were trained in advance.

Focus group interviews. To get more in-depth information on differences between trainers and trainees in how they implemented and experienced the intervention program including professional development activities, we conducted semi-structured focus group interviews after the intervention period with all participating teachers at the schools where we observed during the intervention period. Hence, we conducted five interviews at five different schools, with 9 trainers and 11 trainees in total (unfortunately, 1 trainer and 1 trainee were not able to participate). The interview protocol focused on three main themes: the lesson program Tekster, the teacher manual and the training sessions. The results from the quantitative data were used as input for these themes, in order to further explain differences between trainers and trainees and to explore possible factors that might have affected intervention fidelity.

The protocol began by asking trainers about their experiences with the implementation of the program in their classroom, differentiation between weak and strong writers, the content of the lessons, and the general format of the lessons, i.e., the steps of the writing strategy. We used open, non-directive questions to limit socially desirable responses and followed up with 'why' and 'how' clarification questions until we reached full understanding. After that, we asked trainees the same questions, using the same procedure. When both trainers and trainees did not have anything to add to this theme, we continued with the second theme, the teacher manual. For this theme, we asked teachers to indicate whether they used the general introduction, the lesson plans, and the benchmark rating scale in the teacher manual for preparing their writing lessons, and what their opinion was on these aspects of the manual. We applied the same procedure: first we asked the trainers, then the trainees. The same procedure was followed for the third theme, the training sessions, in which we asked both trainers and trainees whether they experienced enough support to execute the lessons. To get more information on the content of the teacher-training sessions, we asked trainers what aspects of the training program they transferred to their colleagues and whether they experienced enough support to do this. Trainees were asked to indicate how often they experienced some sort of support from their trainers and what kind of support that was. Finally, teachers were offered the possibility to offer recommendations for improving the lesson program Tekster based upon their experiences.

Interviews were conducted by two interviewers: one asking the main questions, the other taking notes and summarizing what was said during the interview in order to enhance the reliability and validity of the interpretation of the interview data. The interviews were audiotaped and were 50 to 67 minutes in duration. Afterwards they were transcribed and coded by an independent coder, an undergraduate student who was trained in advance on the core features of the intervention program. For each interview separately, (sub)categories were identified and coded. Next, the subcategories of all interviews were compared in order to identify important themes and subcategories of information across the five interviews. The themes that were identified from the transcripts were comparable to the themes and subcategories from the interview protocol. We also marked specific quotes that served as a clear illustration of the views of trainers and trainees.

Data analysis

Students' writing quality. Scores for students' text quality were hierarchically organized; scores were cross-classified with students and tasks, and students were nested within classes. The data are therefore analyzed by applying six multilevel models in which parameters were added systematically, in order to test the effectiveness of the intervention program and to test whether there are differences between the trainer and trainee group. In this type of models all students, including those with partly missing values, are taken into account. Inspection of the data revealed that only 4% of the students missed four or more of the writing tasks.

Model 1 is the basic null model in which we only account for random error (S^2_e) and random effects of students (S^2_s), tasks (S^2_t), and classes (S^2_c). That is, writing scores are allowed to vary within and between students, between tasks (including systematic variation due to genre), and between classes. In Model 2 training group is added as a fixed effect to test whether average scores differ between trainers and trainees. In Model 3 measurement occasion is added as a fixed effect to test whether average scores differ over time. Model 4 tests the main effect of the writing intervention by estimating the interaction between group and measurement occasion. As trainers and trainees received the intervention at different time intervals, we tested the interaction by comparing the slope of the regression line of students' writing scores in the trainer group between the first and second measurement occasion with the slope of the regression line of students' writing scores in the trainee group between the second and third measurement occasion (instead of comparing average scores on the pre- and post measures). This model includes the restriction that the effect of the intervention is the same in the two groups. In Model 5 this restriction is removed to test whether the effectiveness of the intervention differs between the trainer and trainee group. In Model 6 grade is added as a fixed effect to test whether average scores differ between grades.

Teachers' self-efficacy and attitudes for writing. To analyze how the intervention including the professional development program affected both trainers and trainees, we analyzed differences between teacher-trainers and teacher-trainees on self-efficacy, and attitudes for writing and teaching writing. We used a repeated measures MANOVA to determine whether teachers' feelings of self-efficacy and attitudes for writing and teaching writing changed over time, due to following the intervention program. We also tested whether there were differences between trainers and trainees, and whether there were differences between grades.

Implementation and social validity of the intervention. Teachers' implementation of the intervention program was measured by logbook entries after each lesson, questions in the post-intervention questionnaire, and classroom observations. We first analyzed these measures separately. For each measure we analyzed whether there were significant differences between trainers and trainees in how they implemented the lesson program, or key components of the program. Next, we triangulated the quantitative data in order to check whether they converge or diverge. In a following step, we checked whether the interview data confirmed the quantitative data and whether they explained the most prominent findings from the quantitative measures. We also used teachers' quotes from the interview data to illustrate quantitative findings.

RESULTS

Effect of the intervention on students' writing quality

We first tested the effect of the intervention on students' writing performance. Table 9.3 shows the results of the fit and comparison of the planned models. As can be seen, there was no main effect of training group (Model 2 versus Model 1, $\chi^2(1) = 1.74$, $p = .19$), indicating that average writing scores did not differ between students from teachers who were trained by experts and students from teachers who were trained by their colleagues. There was, however, an effect of measurement occasion (Model 3 versus Model 1, $\chi^2(3) = 101.61$, $p < .001$), indicating that scores were not the same on the three measurement occasions. Results also showed an interaction effect between training group and measurement occasion (Model 4 versus Model 3, $\chi^2(2) = 4564.73$, p < .001), indicating that differences in writing scores between two measurement occasions (i.e., between occasion one and two, or between occasion two and three) were not the same for students in the trainer and trainee group.

Table 9.3 Fit and comparison of nested models

			Comparison			
Model	N_{pars}	-2 Log Likelihood	Models	ΔX^2	Δdf	*p*
1 basic null model	5	83570.16				
2 + training group	6	83568.42	2 vs 1	1.74	1	0.19
3 + measurement occasion	8	83468.55	3 vs 1	101.61	3	< .001
4 + training*measurement occasion	9	78903.82	4 vs 3	4564.73	2	< .001
5 + training*measurement occasion (differential effects of intervention)	10	78902.32	5 vs 4	1.50	1	.22
6 + grade	12	78870.53	6 vs 4	31.79	2	< .001

With the fifth model we tested whether the effectiveness of the intervention (indicated by an interaction between training group and measurement occasion) was different between the trainer and trainee group. This model was not significant (Model 5 versus Model 4, $\chi^2(1) = 1.50$, p = .22), indicating that there were no differences between trainers and trainees in the effectiveness of the intervention.

Table 9.4 Average writing scores and their variances for students in the two groups; standard errors in parentheses

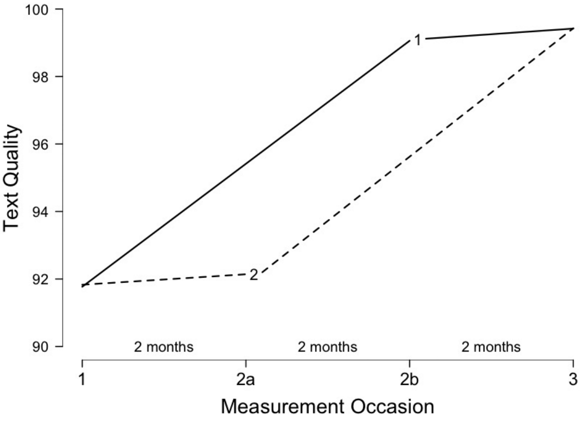
	Parameter	*SE*	*t*	*p*
Fixed part				
Trainer group	91.77	1.12	81.94	<.001
Δt2	+7.29	0.58	12.57	<.001
Δt3	+0.36	0.64	0.56	.29
Trainee group	91.83	1.07	85.82	<.001
Δt2	+0.31	0.60	0.52	.30
Δt3	+7.29	0.58	12.57	<.001
Random part	S^2	*SE*		
Classes	32.57	6.43		
Tasks	15.66	1.26		
Students	50.45	2.47		
Error	78.44	1.19		

Parameter estimates of the fourth model are summarized in Table 9.4. While generalizing over students, teachers, and tasks, students' writing improved with 7.29 points ($SE = 0.58$, $t = 12.57$, $p < .001$), which was a medium effect (Cohen's $d = 0.55$). Between the first two measurement occasions, students in the trainee group served as a control group. It can be seen that their writing performance did not improve during this period ($\beta = 0.31$, $SE = 0.60$, $t = 0.52$, $p = .30$). Further, the effect of the intervention maintained over time: there were no significant differences in writing scores between the posttest directly after the intervention and the delayed posttest after two months in which students in the trainer group returned to their regular writing program ($\beta = 0.36$, $SE = 0.64$, $t = 0.56$, $p <$ .29). A graphical overview of the effect of the intervention in both conditions is presented in Figure 9.1.

Table 9.3 further shows a main effect of grade (Model 6 versus Model 4, $\chi^2(2) = 33$, $p < .001$), indicating that average writing scores were different for students in grade 4 to 6. On average, grade 5 students scored 5.65 points ($SE = 1.39$) higher than students in grade 4. The scores of students in grade 6 were even higher: they scored 8.69 points ($SE = 1.39$) higher than the students in grade 4. Hence, the average improvement in writing quality per grade yielded 4.78 points. If we compare the improvement due to the intervention to the general improvement in writing skills of students between grade 4 to 6, the magnitude of the effect of the intervention becomes even more visible. This comparison shows that students' writing improved by more than one-and-a-half grade after following the four-month writing intervention program.

We approximated the magnitude of the effect of the intervention on students' writing performance, for the whole intervention program consisting of a total of 16 lessons. Inspection of students' workbooks revealed, however, that not all teachers managed to complete the whole intervention program. On average,

Figure 9.1 The effect of the intervention is estimated by comparing the slopes of the regression lines for text quality scores. Solid lines represent scores of the trainer group (1), who received the intervention between measurement occasion 1 and 2b. Dashed lines represent scores of the trainee group (2), who received the intervention between measurement occasion 2a and 3.



they completed 13 lessons (*SD* = 2.80). This affected students' writing scores: for each missing lesson, the effect of the intervention decreased with 0.58 points (*SE* = 0.12, *t* = 5.17, *p* < .001). There was no significant difference between the trainer and trainee group ($\chi^2(1) = 0.43$, *p* = .51).

Effect of the intervention on teachers' self-efficacy and attitudes

The significant improvements in students' writing performance after the intervention indicated that trainers and trainees effectively implemented the writing program in their classrooms. To further analyze how the intervention including the professional development program affected teachers, trainers as well as trainees, we analyzed differences between teacher-trainers and teacher-trainees on self-efficacy and attitudes for teaching writing. Multivariate test results of the repeated measures MANOVA indicate that for at least one of the dependent variables there was a significant main effect of training group (*F*(4, 49) = 3.31, *p* < .05, eta^2 = .21), and time (*F*(4, 49) = 4.05, *p* < .01, eta^2 = .25). There was no main effect of grade (*F*(12, 129.93) = 0.93, *p* = .52). Table 9.5 shows the average scores for teachers' self-efficacy and attitudes toward writing and teaching writing over time. Univariate test results indicate that teachers were more positive about teaching writing after the intervention than before (*F*(1, 52) = 5.99, *p* < .05, eta^2 = .10), they felt more efficacious to teach successful writing (*F*(1, 52) = 8.10, *p* < .01, eta^2 = .14), and they felt more efficacious to support inexperienced writers (*F*(1, 52) = 10.52, *p* < .01, eta^2 = .17). Their general attitudes towards writing remained the same (*F*(1, 52) = 1.15, *p* = .29). Although the average scores for teachers' self-efficacy for successful writing and their attitudes towards teaching writing seemed to be slightly higher for trainers than for trainees (respectively, $F(1, 52) = 8.11, p < .01, eta^2 = .14$ and $F(1, 52) = 8.42, p < .01, eta^2 = .14$), there were no significant interaction effects between time and group for any of the subscales ($F(1, 52) < 1.56, p > .22$). Taken together, the results indicate that the intervention had the same effect on trainers and trainees: both groups of teachers became more positive and felt more efficacious about the teaching of writing.

Table 9.5 Means and standard deviations of teachers' self-efficacy and attitudes towards writing and teaching writing

	Before	After
	M (*SD*)	M (*SD*)
Attitudes towards writing	3.13 (0.80)	3.05 (0.88)
Attitudes towards writing instruction	3.70 (0.73)	3.93 (0.55)*
Efficacy for teaching successful writing	3.30 (0.69)	3.56 (0.62)*
Efficacy for supporting struggling writers	3.55 (0.50)	3.72 (0.46)*

Note. * indicates that means were significantly different over time, with F (1, 52) > 5.99 and p < .05.

Teachers' classroom implementation of the intervention

Implementation of the lesson program. The logbook and questionnaire data, together with the classroom observations provided information on how teachers implemented the lessons in the classrooms. In total, 87% of the logbooks were filled in and returned. The logbook data show that the average preparation time for the lessons was 12 minutes ($SD = 6.4$), which did not differ between trainers and trainees ($F(1, 93) = 1.28, p = .26$). The average lesson time as reported by teachers in the logbook was 46 minutes ($SD = 6.98$), which was according to the planning in the lesson plans in the teacher manual and did not differ significantly between trainers and trainees ($F(1, 93) = 3.28, p = .07$). The lesson time reported by the teachers converged with the observation data, which showed that the duration of the average observed Tekster-lesson was 43 minutes. However, not all lessons were fully completed in this time: in 28% of the grade 5 and 6 lessons the post-writing phase (reread, evaluate and revise) was not covered during the observed lessons, due to time constraints or differences in pace between students. Teachers indicated that these components would be accomplished later in the week, during students' independent work time. The observational data also revealed that teachers were on task on average 90% ($SD = 10$) of the lesson time, which means that they were engaging in the activities as prescribed in the lesson plan, either plenary with all students (especially during the first phases of the lessons), or with individual students. There were no significant difference between trainers and trainees in on-task behavior ($F(1, 52) = 0.70, p = .41$). Questionnaire data revealed that most of the teachers (86%) adjusted some or all lessons to their own context. There were no differences between trainers and trainees ($\chi^2(2) = 0.92$, $p = .63$). Main reasons for adjustments were: classrooms combining multiple grades (combining lessons from different workbooks), students with Dutch as L2 (additional language support), or including examples from real-life, the news or students' own experiences.

Together, the quantitative findings showed that teachers implemented the content and structure of the lessons as intended, but they struggled to finish all lessons in the planned time. The qualitative findings from the interviews confirmed these quantitative findings. During the interviews, both trainers and trainees indicated that for a number of lessons the planning of the lessons was too tight, and that they did not succeed in completing all lesson components during the planned lesson time. This was especially the case for grade 5 and 6 teachers, for whom the lessons included more post-writing activities. Not all students were able to finish their texts at the same time, which complicated the organization of post-writing activities. As suggested in the manual, most of the teachers divided the lessons in two parts. In the first part students planned and organized ideas and started with writing their text. In the second part students evaluated and revised their text. Between the first and second part students were able to finish their text at their own pace.

A very positive organizational point according to the teachers was that they only needed little time to prepare the lessons: only 10 minutes on average as indicated by the questionnaire data. In the interviews it became clear that most teachers spent even less than 10 minutes on preparing their lessons. They indicated that the lesson plans and the goals of each lesson in the manual were very clear. They further indicated that preparation was facilitated by the well-defined overall structure of the lessons: for each lesson, regardless writing genre, students had to apply the same steps of the writing strategy. Some teachers reported that they adjusted the content of writing lessons to fit students context and/or experience even more. For example, in one of lessons students had to write an invitation for the end-of-year musical. One of the teachers had her students write an invitation for their own musical, based on this writing task. Another teacher collected recent examples from the news and internet for text structure instruction instead of the examples provided in the lesson material. Teachers from multiple grade classrooms indicated that during their lesson preparation they paid specific attention to how to combine the lessons for students in different grades into one writing lesson for the whole class. This was especially challenging for lessons in which the communicative goal of the lesson was not the same across grades. Furthermore, one trainee indicated that the preparation depended on her familiarity with the text genre: For more unfamiliar genres, she had to collect additional background information and example texts to increase her own knowledge about the particular genre.

Implementation of key components. Regarding the implementation of key components of the intervention program, i.e., modeling, strategy use, feedback, and evaluating text quality, questionnaire data showed that all teachers modeled (a part of) the writing process during the intervention: 57% of the teachers modeled in some lessons, 35% of the teachers modeled in every lesson, and 8% of the teachers indicated to model multiple times in every lesson. One-third indicated that this was more frequent than during their regular writing instruction, half of them indicated that it was the same. For feedback a similar result was found: 62% of the teachers indicated to have provided feedback in some of the lessons, 27% provided feedback in every lesson, 10% provided feedback multiple times in every lesson, and only 2% provided feedback only once. Again, one-third of the teachers indicated that this was more frequent than in their regular practice, and half them indicated that it was the same. The majority of the teachers provided both process- and product-related feedback, and most teachers provided a combination of oral and written feedback. There were no differences between trainers and trainees in how often they implemented modeling and feedback in their lessons, $\chi^2 < 4.39$, $p > .11$.

The questionnaire data converged with the observational data, which indicated that both trainers and trainees implemented the key components of the intervention in their classrooms to an equal extent. On average, they modeled once during each lesson ($M = 1.00$, $SD = 1.32$). On average, they explicated the acronym once ($M = 1.22$, $SD = 1.20$), and the strategy twice ($M = 2.13$, $SD = 1.55$) during the lessons. Feedback was provided multiple times ($M = 2.76$, $SD = 1.54$). The results of the observational data suggest that teachers executed the lessons as planned in the lesson plans in the manual and there were no differences between trainers and trainees in the observed components, $F(1, 52) < 2.80$, $\mathrm{p} > .10$. There were, however, differences between trainers and trainees, in how often they evaluated their students' written products ($\chi^2(4) = 16.46$, $p < .01$). The questionnaire data showed that the majority of the trainers did this only once (63%) or in some lessons (35%), whereas the majority of trainees evaluated students' texts in some lessons (47%) or in every lesson (29%). Whereas 21% of the trainers indicated that they evaluated text quality more frequently than in their regular practice, this was only found for 9% of the trainees ($\chi^2(2) = 6.07$, $p < .05$). Results also showed that trainers and trainees differed in the use of the benchmark rating scale for evaluating text quality (93% versus 74%, $\chi^2(1) = 4.16$, $p < .05$) or providing feedback (86% versus 62%, $\chi^2(1) = 4.75$, $p < .05$). Forty-one percent of the teachers also indicated to use the benchmark rating scale for writing instruction in class, which did not differ between trainers and trainees ($\chi^2(1) = 0.28$, $p = .60$).

During the interviews, trainers and trainees reported hardly any differences in how they implemented the key components of Tekster. They indicated that the writing strategy changed the content of their lessons. Because the lessons were divided into steps of the writing process, teachers focused more on the process of writing, instead of mainly on the product. Whereas the focus of instruction changed, it became clear during the interviews that teachers struggled with adapting their mode of instruction. Teachers who were used to apply modeling during instruction applied this also in their writing lessons. However, some of them indicated that they intentionally did not model, as they were afraid that students would not be able to come up with information themselves. Teachers who were not used to apply modeling, indicated that they would like to have had more training and practice in teacher modeling, as they found it sometimes hard to decide what to model and how to do it. One of the trainers who was already experienced in modeling, indicated that the training offered enough support to implement modeling effectively in the writing lessons and to instruct a colleague, but that this might pose a bigger challenge for teachers who are not yet familiar with the underlying principles of modeling. Moreover, when teachers provided support to students, for instance by modeling or feedback, this was mostly offered during the prewriting and writing phase. During the post-writing phase students worked mainly independently without extra teacher support. According to the teachers, this was due to lack of time. This was also the main reason for not evaluating all students' final writing products or providing feedback on a regular basis.

Implementation of the teacher training. From the post-intervention questionnaire it became apparent that the collegial training sessions differed from the expert training sessions, both in terms of duration as well as in the content that was discussed during the sessions. The two expert sessions lasted 8 hours in total, whereas the collegial training session consisted of two or multiple sessions of, on average, 93 minutes in total. However, the duration of the collegial training differed largely between teachers ($SD = 83$), and ranged between 0 and 480 minutes. For most of the teachers (71%) the training sessions were one-on-one. There were only four trainees who attended a Tekster-lesson of their colleague. The topics that were discussed during the collegial training sessions were the goal and structure of the lesson program (84%), organisation of the lessons (84%), feedback (81%), rating text quality (78%), modeling (76%), benchmark rating scale (73%), student texts (46%), and specific problems (46%).

That the content of the professional development program substantially differed for trainers and trainees became also clear from interviews with teachers. Teachers confirmed that the collegial training sessions were relatively short and mainly dealt with the content of the intervention and organizational issues. Moreover, trainees did not study the provided background information in the teacher manual. As a result, trainees were less aware of the importance of the topics that were covered during the professional development program, such as modeling, feedback and assessment of text quality. This can partly explain why trainees struggled with changing their mode of instruction. It is quite remarkable that trainers did not transfer their newly acquired knowledge about effective instructional modes, as trainers reported that they found the information concerning modeling, feedback and assessment of text quality particularly useful. For instance, one trainer stated that benchmarks helped students to "understand what constitutes a good text", which motivated students to revise their own texts.

A possible explanation why trainers did not address the mode of instruction is perhaps that they overestimated their colleagues. For instance, one trainer stated that she did not put a lot of effort in the training session with her colleague, as "she is a very experienced teacher". The trainees themselves indicate that they hardly needed any support from their trained colleague, as "the teaching materials and the manual were very clear".

Social validity

The acceptability of and satisfaction with the intervention was established by combining logbook data, questionnaire and interview data on teachers' experiences with the lessons, the key components of the program, the manual and the training sessions.

Teachers' experiences with the lesson program. With regards to the lesson program, logbook data revealed that, on average, teachers estimated the level of difficulty of the lessons for their students with 3.03 ($SD = 0.47$) on a five-point scale ranging from 1 (easy) to 5 (hard). This indicated that the lessons were challenging, but not too difficult for the students. Further, teachers reported on the level of difficulty of executing the lessons, also on a five-point scale ranging from 1 (easy) to 5 (hard). This aspect was rated with an average score of 2.39 ($SD = 0.50$), indicating that teachers felt sufficiently equipped to teach the lessons. For these variables, there were no differences found between trainers and trainees ($F <$ 3.28, $p > .07$). Teachers generally were positive about the teaching program. In the logbooks, trainers were slightly more positive than trainees (respectively M = 7.93, $SD = 0.49$ and $M = 7.70$ on a 10-point scale, $SD = 0.52$, $F(1,93) = 5.58$, $p = .02$). The post-intervention questionnaire showed that trainers and trainees were equally positive about the program as a whole ($M = 4.10$ on a 5-point scale, $SD = .62$, $F(1, 62) = 3.14$, $p = .08$). Their students were also positive about the lessons ($M = 6.75$ on a 10-point scale, $SD = 2.23$).

The interview data confirm the quantitative data. Especially the communicative goals of the writing tasks during the lessons were frequently mentioned as a positive aspect of the program. In contrast to the writing tasks from the language textbooks that they used to work with, Tekster offered them writing tasks with clear communicative goals that were close to the experience and interest of their students. Trainers as well as trainees indicated that these kind of writing tasks helped them to focus more on the content of students' writing instead of on formal aspects, like mechanics and conventions, during instruction. Moreover, because the writing tasks were close to students' own experience and the content of their writing was taken more seriously by the teacher, students seemed to be more motivated to write. However, this motivation was not always evident. Both trainers and trainees from different schools indicated that students sometimes started demotivated, "bleeeh, Tekster again!". This changed to a more positive attitude during writing, especially when they were working together with peers.

Table 9.6 Means and standard deviations of teachers' satisfaction with the intervention

	Trainers *M* (*SD*)	Trainees *M* (*SD*)	Total *M* (*SD*)
Tekster lesson program			
General lesson program	4.24 (0.58)	3.97 (0.63)	4.10 (0.62)
Strategy instruction	4.59 (0.57)	4.44 (0.75)	4.51 (0.67)
Teacher modeling	4.41 (0.57)	4.24 (0.74)	4.32 (0.67)
Peer modeling in video-clips	2.83 (0.89)	3.06 (1.13)	2.95 (1.02)
Discussing model texts	4.31 (0.66)	4.15 (0.74)	4.22 (0.71)
Peer interaction	3.76 (0.87)	3.74 (0.99)	3.75 (0.93)
Feedback	4.17 (0.62)	3.91 (0.67)	4.03 (0.65)
Teacher manual			
Providing lessons	4.24 (0.87)	4.03 (0.87)	4.13 (0.87)
Using benchmark rating scale	4.62 (0.68)*	4.00 (1.16)	4.29 (1.01)
Training sessions			
Providing lessons	4.21 (0.77)*	3.06 (1.23)	3.59 (1.19)
Modeling	3.66 (0.90)*	2.79 (0.95)	3.21 (1.01)
Providing feedback	4.17 (0.82)*	2.59 (0.93)	3.32 (1.18)
Evaluating text quality	4.31 (0.71)*	2.97 (1.11)	3.59 (1.16)

Note. * indicates that means between trainers and trainees were significantly different, with F (1, 62) > 6.47 and p < .05; for all the other components means between trainers and trainees were not significantly different, with F (1, 62) < 3.14 and p > .08.

Teachers' experiences with the key components. Teachers indicated in the questionnaire that the key components of the writing program such as the strategy, modeling, model texts, feedback, and peer interaction, supported their students' writing performance, see also Table 9.6. We found no differences between trainers and trainees. Moreover, both trainers and trainees indicated that they intended to keep using components of the intervention program, such as modeling (84%), strategy-instruction (84%), the acronym (81%), discussing model texts (70%), peer interaction (78%), feedback (92%), and specific writing tasks from the lesson program (70%). The majority of trainers did also indicate that they intended to continue using the benchmark rating scales for evaluating text quality (89%), whereas for trainees this was only 50%. Peer modeling using video clips was the only component of which only a minority of the teachers (14%) intended to continue using it.

The interview data made clear that teachers were positive about all the key components of the program. According to them, the writing strategy was the most effective component. Both trainers and trainees indicated that students experienced much support from the steps of the writing strategy. For instance, according to one teacher the benefit of the program is that "students are taken through the process of writing a text, step by step". Another teacher compared the steps to “a recipe that always results in a delicious cake”. Teachers indicate that especially struggling writers experienced a lot of support by the strategy, because “it helped them to commence with writing more easily”. It also turned out that the overall strategy promoted transfer: students continued to use the steps for writing tasks also after the intervention period. The support that teachers experienced from the writing strategy might also explain the improvement in their self-efficacy for teaching writing. One teacher reported that “because you are so focused on the writing process of your students, you start reflecting on your own writing process as well”.

Teachers were less positive about the videoclips in which peers modeled (parts of) the writing process or the writing product, and about the peer-feedback component that was included. They indicated that the quality of the videoclips was insufficient. Some teachers reported that they stopped using the videos, but incorporated this information in the plenary instruction or in a classroom discussion. With regards to peer-feedback, both trainers and trainees indicated that their students lack the skills and knowledge to provide effective feedback on each other’s texts. They revealed that when students provide feedback, they focus mainly on positive aspects in a very general way, such as “well done”, or they provide suggestions which are related to lower order aspects of the text, such as “add a title”, or to formal aspects such as “write more neatly”. Both trainers and trainees indicated that students struggled with providing feedback on the content or communicative goal of the text.

Teachers experiences with the teacher manual and the training sessions. Table 9.6 presents the questionnaire data on teachers’ satisfaction with the professional development activities. In general, teachers were highly satisfied with the teacher manual as well as with the training sessions, as they indicated that both supported them in the execution of the writing lessons. More specifically, the training sessions supported their practice in modeling, giving feedback and evaluating the quality of their students’ texts. There was, however, a difference between teacher-trainers and teacher-trainees: trainers experienced generally more support from the expert training sessions than trainees did from the collegial training sessions ($F(1, 62) > 6.47$ and $p < .05$).

It is important that skills and knowledge on how to implement key features of a new program are easily transferable between teachers as in elementary education there often are changes in staffing, due to the large number of part-time working teachers and personal circumstances such as maternity leave or sick leave, which was confirmed by the interviewed teachers. In the interviews trainers indicated that it was easy to transfer the content of the training sessions to their colleagues and trainees reported that they experienced enough support from the trainers and the teacher manual to execute the Tekster-lessons. Despite this, the interviewed teachers also indicated that it often takes a couple of years to implement a program like this in the most optimal way. One of the trainers stated that “the longer you work with it, the better it gets”.

DISCUSSION

In this study, we examined the effect of the writing intervention program Tekster, including professional development activities for teachers, on students' writing performance and on teachers' self-efficacy and attitudes for writing and the teaching of writing. Tekster is a comprehensive program for writing for the upper elementary grades, combining strategy-instruction, text instruction, and the teaching of self-regulation skills with observational learning, explicit instruction, and (guided) practice to address both the focus and mode of instruction. Professional development of teachers was promoted by means of a teacher manual including lessons plans and training on how to optimally implement key components of the intervention in the classroom. To investigate whether the content of such a professional development program could be transferred between teachers, we applied a teachers-training-teachers approach in which half of the participating teachers (trainers) were trained by the researchers, and these trainers subsequently trained one or more colleagues (trainees).

The quasi-experimental results showed that Tekster improved students' writing quality significantly (ES = 0.55). After the intervention, students wrote better texts than students who were engaged in their regular writing activities. As Tekster was tested on a large scale involving 1365 students and 68 teachers from 25 schools, and students' writing performance was assessed with nine writing tasks in three genres, this effect is generalizable over students, teachers, and tasks. Moreover, the effect of the program maintained over time: two months after the intervention, students still wrote qualitative better texts than they did before the intervention. Results further show that there were no differences between the writing performance of students of teacher-trainers and teacher-trainees, indicating that trainers and trainees implemented the intervention with equal effectiveness. Results also show that both trainers and trainees became more positive and felt more efficacious about teaching writing after the intervention. Their general attitude toward writing was not changed by the intervention program.

As teachers had to implement the intervention program in their own classrooms, we examined whether they implemented the program and the collegial training sessions with fidelity. We also investigated the social validity of the intervention program and the teachers-training-teachers approach, as this provides valuable information about the feasibility of implementation in classroom practice. We used an explanatory sequential mixed methods approach in which we collected and analyzed quantitative data from post-intervention questionnaires, logbooks, and classroom observations, followed up by qualitative data from focus group interviews.

The mixed methods data revealed that teachers implemented the intervention program as intended and that there were no differences between trainers and trainees in how they implemented the intervention. For trainers as well as for trainees the focus and mode of instruction changed by working with the intervention. They reported to use a more process-oriented approach for teaching writing in which they incorporated modeling. Further, they supported students' self-regulation by explicitly strategy instruction and focusing on communicative goals of writing. They also adapted their instruction to the need of individual students by providing feedback, both on the written product and the writing process.

Concerning the social validity of the intervention, experiences of both trainers and trainees with the lesson program, the teacher manual and the training were highly positive. They especially appreciated the strategy-focused approach in the writing lessons and the writing tasks with explicit communicative goals. In the teachers' perceptions, their students' writing was supported by the program, and they intended to keep using components of the program.

Implications for improving writing education in upper elementary grades

The intervention program Tekster combines several instructional practices that have proven to be effective in earlier research into one comprehensive program for writing. Although we did not measure the effectiveness of individual components, the mixed methods data on teachers' implementation and experiences with the components of the program provide valuable clues on what aspects of the program may have been especially effective for improving students' writing and what aspects may have been less effective. To improve writing education in upper elementary grades, it is important to know what works and what not. Therefore, we will elaborate on these aspects in more detail.

Effectiveness of strategy-instruction. In line with previous research (cf. Graham, 2006; Graham et al., 2012; Koster et al., 2015) the results from the present study show that strategy-instruction is an effective instructional practice to enhance students' writing performance. The application of a writing strategy reduces the number of cognitive processes that are active at the same time, which reduces students' cognitive overload and ultimately leads to improved writing performance (Graham et al., 2012; Kellogg, 1988). The interviewed teachers in this study confirm that the steps of the writing strategy offered students support to manage their writing process more effectively. They indicated that by generating and organizing ideas before writing, students came up with more ideas than before the intervention and that students wrote longer and better texts.

The strategy that was the core of every Tekster-lesson also elicited a change in teachers' classroom practice: the emphasis of teachers' instruction shifted from a product to process approach to writing. Teachers provided more support to students during the writing process, they modeled parts of the writing process, and provided feedback during the process. Interview data also revealed that teachers were better able to differentiate their instruction to meet the needs of weak as well as proficient writers. This is confirmed by their self-reported feelings of efficacy for teaching writing: both trainers and trainees indicated that they felt better equipped to teach strong and weak students to write. Although teachers paid more attention to the writing process in general, the results suggest that this was especially true for the prewriting phase, and less so for the post-writing phase. In interviews teachers indicated that students often had not finished their text during regular lesson time, which meant that the final steps of the strategy were finalized during independent work time, with considerably less support of the teacher. This indicates that the post-writing phase is not implemented in the most optimal way. Further research should investigate how the post-writing phase can be implemented more effectively.

Effectiveness of the communicative goals for writing and benchmark scales. The results from this study also emphasize the importance of explicating communicative goals for writing, as well as writing tasks that are close to students' own experiences. Teachers stated that because the lessons were meaningful for students they were willing to put effort in writing their text, even when they were not motivated to write at the start of the lesson. In previous research it was established that setting goals for writing increases self-regulatory skills in writing (Zimmerman & Risemberg, 1997). The results from the present study seem to be in line with this. Teachers indicated that students had a better understanding of why they were writing a text and for whom, and were more able to monitor progress towards these goals, which might explain the improvements in text quality.

The mixed methods data also reveal that the explicitly stated writing goals supported teachers' instructional practice. Teachers were more aware of the goal of writing, which facilitated providing feedback on students' writing products. Teachers' feedback practices were further supported by the benchmark rating scales that were included in the intervention program. In the interviews, teachers indicated that the example texts in the rating scale, which reflect the range of students' writing performance for each writing genre, helped them to evaluate the communicative effectiveness of students' written texts. The data from questionnaires and interviews revealed that teachers used these scales not only for evaluating the quality of students' texts, and providing feedback, but also for classroom instruction. However, there were differences between trainers and trainees in both the implementation and satisfaction with the benchmark scales. In general, trainers were more satisfied with the benchmark scales and they used these scales more frequently than trainees did. This suggests that specific training on how to use these benchmark rating scales is essential for an optimal implementation in the classroom.

Effectiveness of peer interaction. During Tekster lessons, peer interaction is implemented at multiple occasions: students interact with each other before they start writing, e.g., to jointly generate ideas, but also after writing in order to evaluate texts written by their peers. The experiences of teachers with this component of the lesson program were mixed. On the one hand, teachers indicated that interaction between peers enhanced students' motivation to write because students experience that their text is meant for a reader, and that their text is actually read. On the other hand, whereas it was the aim of peer interaction to raise awareness of the effect of their text on a reader, teachers indicated that students found it difficult to give their peers adequate feedback. According to the teachers, students generally responded positively to their peers, and when they did provide critical feedback this was mostly directed on lower order aspects in the text, such as bad handwriting, and errors in spelling or punctuation. Teachers stated that students need more support in order to provide effective feedback to their peers. This can, for instance, be done by including explicit, directive questions in the evaluation step of the writing strategy focusing on the information about text structures and criteria for a good text that were discussed during the introduction of each lesson. An important role for the teacher during this phase is to model to provide feedback and show what effective feedback looks like. Thus, it is essential that teachers not only guide and support the prewriting and writing phase, but also offer support during the post-writing phase.

Previous research showed that teachers themselves also struggle with providing effective feedback that is adjusted to the needs of the students (Bouwer et al., 2016b). Although we focused on how to provide effective feedback during the teacher training sessions, it was beyond the scope of this study to determine whether teachers actually provided more effective feedback to students' writing. Further research is needed to investigate specifically whether teacher training enhances teachers' feedback on students' texts.

Effectiveness of observational learning. To effectively improve writing education, both the focus and mode of instruction should be addressed (Hillocks, 1984; Graham et al., 2012, Koster et al., 2015, Rijlaarsdam & Couzijn, 2000). The focus group interviews revealed that teachers struggled to adapt the mode of instruction, and that they especially encountered difficulties with the implementation of observational learning during writing instruction. Observational learning is an important component in the mode of instruction of Tekster, and it is implemented in two ways: through teachers who model parts of the writing process in class and through video clips in which peers model how they write texts in specific genres. Previous research indicated that through observing a model performing (part of) a writing task while thinking aloud, students gain insight into the writing process. Moreover, it provides students with the opportunity to direct their full attention to the learning task as the learning-to-write is separated from text production (Graham et al., 2005; Rijlaarsdam & Couzijn, 2000; Zimmerman & Risemberg, 1977).

With regards to teacher modeling, the results from this study showed that teachers implemented modeling in their writing, especially during the prewriting and writing phase. Modeling was seldom applied during the post-writing phase: students often had to finish this last step independently. Our results further indicate that some teachers struggled to implement modeling effectively in their instruction. Teachers who were already experienced in modeling stated during the interviews that effective modeling needs extensive practice and training, and that the training that was offered to them may not have been sufficient for teachers who have less experience in modeling. The interview data also revealed that some teachers did intentionally not model as they were afraid that it would limit students' creativity. Some teachers indicated that their students literally copied the text that was modeled without coming up with their own ideas. In their opinion, students did not learn from this. However, research on observational learning suggests that even when students copy ideas from the teacher they still learn from observing the writing process (Schunk, 2012; Zimmerman & Risemberg, 1997). This aspect of modeling should be investigated in more depth in further research.

The peer modeling videoclips were the aspect of the program that teachers were least satisfied with. In the interviews and in the logbooks teachers indicated that the quality of the sound of the videos was insufficient. Further, teachers indicated that students found it difficult to identify with the persons in the videos, which might have distracted students from learning from the content of the videos. Thus, our results indicate that peer modeling in this study was not implemented in the most optimal way. Future research should examine into more detail how peer modeling can be operationalized and organized in a way that it promotes students' learning.

Professional development of teachers

The aim of including professional development in the intervention program was twofold: to support teachers in implementing the intervention more effectively and with more fidelity, and to improve writing education on the long term, beyond the intervention (McKeown, Fitzpatrick, & Sandmel, 2014). To induce lasting improvement in the way writing is taught, the skills and knowledge of teachers have to be enhanced. Training teachers in applying effective writing practices increases their feelings of self-efficacy for teaching writing, which is positively related to their quality of instruction (De Smedt et al, 2016). Our findings indicate that teachers' feelings of self-efficacy were enhanced, suggesting that teachers themselves feel better equipped to teach writing after the intervention. The combination of enhanced self-efficacy for teaching writing with improved lesson material might have elicited a change in teachers' classroom practice. Mixed methods data from observations, interviews and questionnaires revealed that they adapted both their focus and mode of instruction. Moreover, these changes seem to last above and beyond the intervention: even when teachers no longer have the Tekster-lessons at their disposal, their students' writing performance does not significantly decrease compared to their posttest performance.

Further, there were no differences between trainers and trainees in how they implemented the intervention, indicating that both teachers who are trained by experts and teachers who are trained by colleagues were capable of effectively implementing a comprehensive intervention program in their daily classroom practice. From these results we can conclude that a professional development program can be transferred between teachers from the same school, as we found no differences between trainers and trainees in student outcomes, in teachers' perceived feelings of self-efficacy, attitudes toward writing instruction, classroom practice and experiences with the program. This is in agreement with prior research on spillover effects of professional development (Borko, 2004; Lieberman & Friedrich, 2007, Sun et al., 2013).

The only significant difference between trainers and trainees concerns the perceived support of the training sessions: trainers experienced in general more support from their expert training sessions than trainees did from their collegial training sessions. That the teachers trained by experts were more positive than their colleagues about the key components of the program, such as modeling, feedback, and assessing text quality, is not surprising, as the interviews revealed that whereas the expert training sessions primarily focused on these components of the intervention, the collegial training sessions primarily addressed organizational aspects of the intervention, such as what components have to be implemented at what time during the lessons. Despite this, the observation data in this study suggest that there were no differences in how teacher-trainers and teacher-trainees implemented the key components during their lessons: both trainers and trainees applied modeling and provided feedback. This suggests that the professional development program contributed to change in teachers' instructional mode to an equal extent. However, during observations was only tallied whether these key components of the intervention were present, and not how they were executed. The quality of applying these key components might differ between trainers and trainees, because interviews revealed that during the collegial training sessions trainees received little support in how to change their mode of instruction. Trainers, on the other hand, reported that these aspects were eye-openers to them during the expert training sessions.

Why is it harder for teachers to adapt the mode of instruction than the focus of instruction? Whereas changing the focus of instruction primarily requires good teaching materials, changing the mode of instruction requires behavioural change, which is a process that takes time, practice and support. As Desimone (2009) indicated, to be effective professional development should be of sufficient duration, both in span of time and in number of hours spent on the activities. Unfortunately, research has not yet established a rule of thumb, but at least 20 hours of contact time spread over a semester is suggested (Desimone, 2009). It might well be that to have more impact on teachers' practice, the professional development activities in this study should have been longer in duration and more intensive. Concerning the transferability of the adaptation of the mode of instruction between colleagues should be noted that effective transfer requires that trainers understand the goals of the professional development program and know how these goals can be achieved (Borko, 2004). The fact that the trainers themselves still struggle with adapting the mode of instruction might be an explanation for the fact that this aspect of the professional development program was not transferred. Future research should therefore especially focus on the role of professional development for the adaptation of the mode of instruction and investigate how professional development should be arranged to effectively address the mode of instruction and under what conditions this aspect is transferable between teachers.

A limitation of this study is that, due to the design of the study, we cannot be certain that the similarities between trainers and trainees can be attributed to a spillover effect, as we did not include a 'material-only' control group of teachers. It might well be possible that only the materials of the program (i.e., the lessons and the manual) have led to improvement in students' achievement. For instance, interview data revealed that teachers experienced much support from the lesson plans in the teacher manual for preparing their lessons. However, teachers indicated that the training supported them in the implementation of the program in their classroom practice, which was reflected in the increase of the scores of perceived self-efficacy.

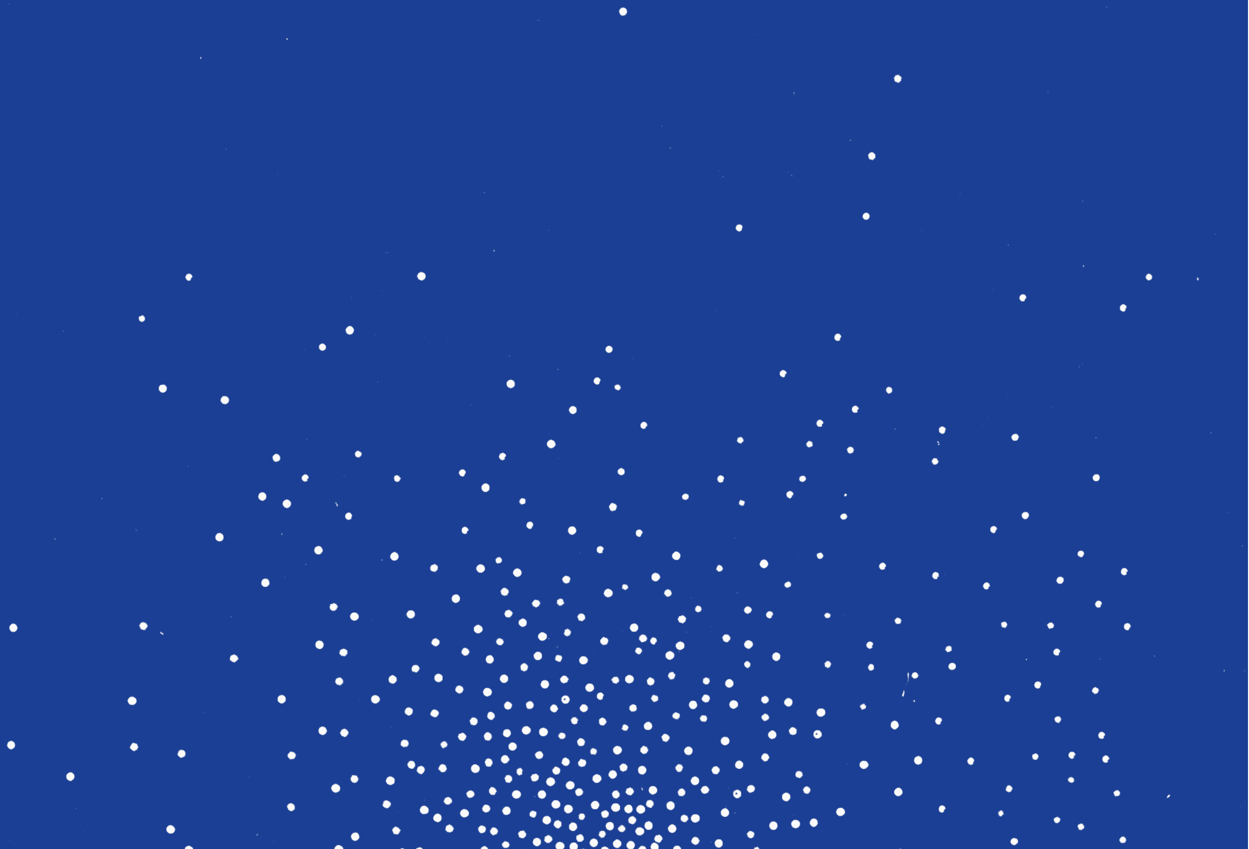
Further, it should be noted that we only tested one form of professional development: a teachers-training-teachers approach with two training sessions with content specific for this intervention. The results of this study can therefore not be generalized and conclusions are limited to this specific intervention study. More research is needed to examine the spillover effect of professional development and the effectiveness of the teachers-training-teachers approach.

Conclusion

In the present study we examined the effectiveness of a writing intervention, Tekster, on a large scale with nearly 70 teachers who implemented the intervention in their own classrooms. We included a professional development program to support teachers' skills and knowledge in working effectively with the program. To make this possible on such a large scale, we applied a teachers-training-teachers approach in which half of the teachers had to transfer their new skills and knowledge to their colleagues. It is shown that Tekster is an effective program for teaching writing and that the content of the additional professional development program is transferable between teachers in the same school: the teachers-training-teachers approach supports teachers to effectively improve their writing education. Not only their instructional focus changed, but it is also a promising approach to change the mode of instruction. All in all, this study provides valuable clues how the gap between research and classroom practice can be bridged to improve writing education.

Chapter 10

GENERAL DISCUSSION

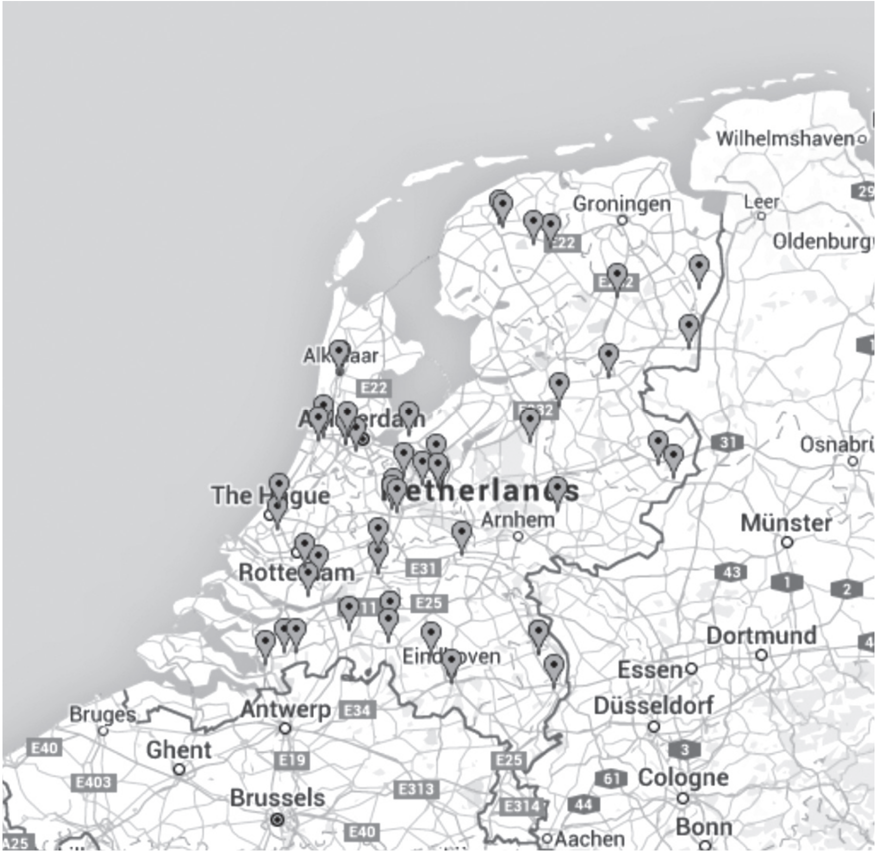


AIM OF THIS DISSERTATION

The aim of this dissertation was to improve writing education in upper elementary grades by developing an effective strategy-focused comprehensive program for writing, including (pre- and post-) writing activities for students, professional development for teachers, and tools for the assessment of writing. Building on the findings of previous research, we developed and tested Tekster [Texter], a comprehensive program for teaching writing in grade 4 to 6. Tekster was developed in close collaboration with elementary teachers. The effectiveness of the program was tested in two large-scale intervention studies in which in total 2766 students and 144 teachers participated, from 125 classes and 52 schools. Figure 10.1 shows how participating schools are spread over the Netherlands.

In this concluding chapter we will summarize the main findings of the studies that were conducted. These findings will be discussed, as well as their implications for future research. We will conclude with a discussion on how the developed writing program, Tekster, can be implemented in upper elementary grades.

Figure 10.1 An overview of the location of participating schools.



SUMMARY OF MAIN FINDINGS

Developing the framework

Chapter 2 to 4 delineate the framework for Tekster. As a starting point in Chapter 2 we examined which instructional practices specifically improve the writing performance of students in grade 4-6 in a general educational setting by means of a meta-analysis. This meta-analysis can be considered as a refinement, an update as well as an expansion of previously conducted meta-analyses (Graham et al., 2012; Graham & Perin, 2007; Hillocks, 1984). This meta-analysis is a refinement in the sense that we only included studies that specifically targeted students in grade 4 to 6 in a regular educational setting. Further, it can be considered as an update of the previous meta-analyses as a quarter of the studies included in our analysis have not been included in prior analyses. Lastly, our analysis can be considered as an expansion of previous meta-analytical research as we not only summarized effect sizes, but statistically compared intervention types to examine whether they differed significantly from each other in effectiveness. The meta-analysis included 32 (quasi-) experimental writing intervention studies which specifically aimed at improving the writing proficiency of students in the upper elementary grades (grade 4-6). We identified the following effective instructional practices: goal-setting (ES = 2.03), strategy instruction (ES = 0.96), feedback (ES = 0.88), text structure instruction (ES = 0.76), and peer assistance (ES = 0.59). Pairwise comparison of these categories revealed that goal-setting is the most effective intervention (although only based on one study comparing several conditions and multiple grades), followed by strategy instruction, text structure instruction, peer assistance and feedback respectively. The results of this study provided us with important clues on what elements to include in a teaching program for writing, aimed to improve students' writing skills.

Understanding what instructional components are important for effective writing education is only part of the puzzle. It is also essential to understand what is the best way to assess writing proficiency, from a research perspective as well as a practical perspective. For researchers it is important to be able to measure writing proficiency in order to make inferences of the effectiveness of a writing intervention, and teachers need to get an impression of their students' writing proficiency and be able to monitor students' progress over time. Therefore, in Chapter 3, we examined how students' writing proficiency can be inferred on the basis of individual writing tasks, and specifically how many texts and raters are necessary in order to determine this. Results showed that raters, tasks, as well as genre account for variance in the measurement of writing, making generalizations to writing proficiency based on one task in one genre, rated by only rater, almost impossible. To draw conclusions about writing proficiency in general, students should at least write three different texts in each of four genres, and these should be rated by at least two raters.

In educational practice, however, it is challenging to organize a writing assessment procedure involving multiple raters. It is therefore of vital importance to examine possible ways to reduce the unreliability of the rater, without restricting raters to such an extent that the validity of their ratings is at risk. We developed a rating procedure in which raters have to assess text quality holistically by comparing each text to five benchmarks that operationalize different levels of writing performance. Examples of benchmark rating scales are included in Appendices I, J, and K. The reliability and validity of a benchmark rating procedure was examined in Chapter 4, by comparing it to a holistic rating procedure without benchmarks and to an analytic rating procedure. From the results it can be concluded that a benchmark rating procedure is a promising approach to assess students' writing performance as it yielded less rater variance than holistic rating procedures, as well as less task-specific variance than analytic rating procedures. Moreover, raters could use one benchmark rating scale to rate writing tasks that differed in topic or genre with the same reliability.

Intervention study 1: Improving students' writing performance with Tekster

Together, the findings from Chapter 2 to 4 provided the basis for Tekster, a teaching program for learning to write for grade 4 to 6. Hillocks (1984) showed that for effectively teaching writing both the focus of instruction (what do we teach?) and the mode of instruction (how do we teach it?) are important. In Tekster, the focus of instruction was targeted by combining the effective instructional practices identified in Chapter 2, i.e., strategy instruction, goal-setting, text structure instruction, peer assistance and feedback. To optimize the teaching of writing, the mode of teaching included observational learning, which separates learning from task performance (Rijlaarsdam, 2005; Schunk, 2012), as well as gradual release of responsibility in which the cognitive load shifts from teacher to student through scaffolding (Pearson & Gallagher, 1983; Wood, Bruner, & Ross, 1976). Additionally, explicit instruction by the teacher was applied to activate background knowledge and to make students aware of the purpose and benefits of the strategy.

The program consisted of three lesson series of 16 lessons, one for each grade level, compiled in a workbook, accompanied by a teacher manual. The program was tested in a large-scale intervention study in a general educational setting, with 76 teachers and 1420 students. This experiment is described in Chapter 5. Teachers delivered the intervention to their own students, over a period of two months (two lessons a week). We used a switching panel design, with two groups and three measurement occasions: the first group worked with the program from the first to the second measurement occasion, while the second group engaged in their regular writing activities. Between the second and third measurement occasion this was the other way round. The findings in Chapter 3 indicated that, in order to draw conclusions about a student's writing proficiency, multiple tasks in multiple genres on different topics should be used. The students participating in our study wrote three texts in different genres, on different topics at each measurement occasion students: a narrative text, a descriptive text and a persuasive letter. Thus, from each student we had nine texts in total. The quality of the students' texts was assessed by three raters, using a benchmark rating scale. We had three different scales, one for each genre, see Appendix K for an example of a benchmark rating scale for persuasive letters.

Findings indicate that students' writing performance improved significantly across all grades and that this improvement was maintained two months after the intervention. Further, the switching panel design allowed us to replicate the results within the study, as similar effects were found in the control group, who received the intervention during the second and third measurement occasion. The effect of the intervention was estimated based on the average number of completed lessons. On average, 10 out of 16 lessons were completed, but this varied considerably between classes. When students completed all 16 lessons, the intervention appeared to be even more effective in improving their writing.

Because the lesson program was delivered by regular teachers, in their own classrooms, we analyzed teacher logbooks and observed in classrooms to get more information on whether they implemented the program as intended. Although it was a challenge for teachers to fit two extra writing lessons into their regular weekly schedule, these fidelity measures revealed that teachers closely adhered to the provided lesson plans and applied the key components of the program: they used modeling in their lessons, frequently referred to the acronym and used the (steps of) the writing strategy for teaching writing. Despite this, results on the level of the student revealed a considerable amount of variance due to classes, indicating that the effectiveness of the intervention depended on who had provided the lessons. However, the intervention appeared to be effective above and beyond differences between classes, as indicated by our multilevel analyses. Moreover, as students' writing proficiency was measured with multiple writing tasks in three different genres it was possible to generalize the results not only over students and classes, but also over tasks. Figure 10.2 provides two example texts of the same student written before and after the intervention, illustrating the improvement in students' writing.

Effects of interventions are often expressed in terms of effect size, which is a measure of the difference between mean scores relative to the variability in the sample (i.e., standard deviation). Due to the design of the study, we have several sources of variation: variance related to class, task, student, and student by task including error variance. If we take all the variance into account, the effect size of the intervention is 0.40. This indicates that the students' writing performance improved by almost half a standard deviation after the intervention, while generalizing over students, tasks, and classes. However, if we would have tested the intervention with only one task and in only one class, the effect size would increase to 0.80.

A closer look at key components of Tekster

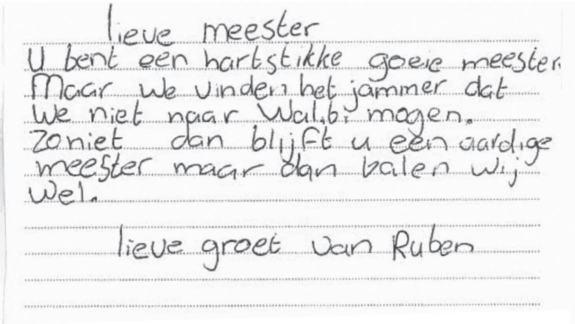
One of the key components of our writing program was a general overall strategy for writing, irrespective of genre, which subdivides the writing process into smaller steps, with the aim to reduce students' cognitive overload during writing.

Figure 10.2 Two examples of a persuasive letter written by the same student, before (left) and after (right) the intervention.

Writing task: Amusement park

You and your classmates want to make a daytrip to an amusement park. Your teacher does not find this a good idea. Still, you want to do everything you can to try to make this happen.

Write a letter to your teacher in which you try to convince him or her with good arguments to go on a daytrip to an amusement park with the whole class. Clearly state in your letter to which amusement park you want to go.



lieve meester
U bent een hartstikke goeie meester.
Maar we vinden het jammer dat
we niet naar Walibi mogen.
Zoniet dan blijft u een aardige
meester maar dan balen wij
wel.

lieve groet van Ruben

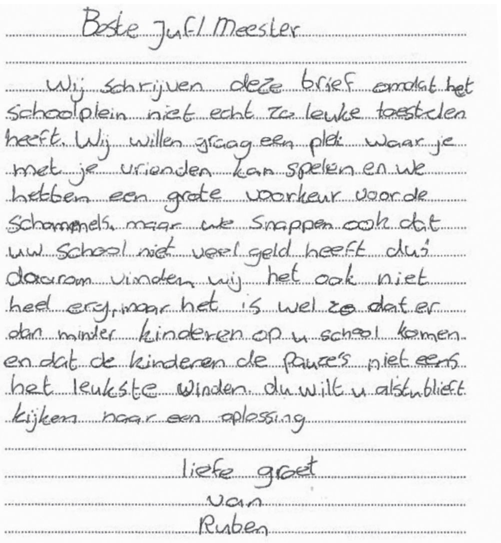
dear teacher
You are a very good teacher.
But we think it's a pity that we are not allowed to go to Walibi.
If not then you will stay a nice teacher but we do think it sucks though.

Dear greeting by Ruben

Writing task: Play equipment in the schoolyard

During the school break you would like to play with your classmates, but the school playground is very boring. Therefore, you have a great idea: new playing equipment in the schoolyard!

Write a letter to your teacher in which you try to convince him or her with good arguments to get new playing equipment in the schoolyard. Clearly state in your letter what this should be (for example, a skate court, climbing frame, marble run) and why it is so important.



Beste Juf/Meester

Wij schrijven deze brief omdat het
schoolplein niet echt zo leuke toestelen
heeft. Wij willen graag een plek waar je
met je vrienden kan spelen en we
hebben een grote voorkeur voor de
Schommels. maar we snappen ook dat
uw school niet veel geld heeft dus
daarom vinden wij het ook niet
heel erg, maar het is wel zo dat er
dan minder kinderen op u school komen.
en dat de kinderen de pauze's niet eens
het leukste winden. du wilt u alstublieft
kijken naar een oplossing

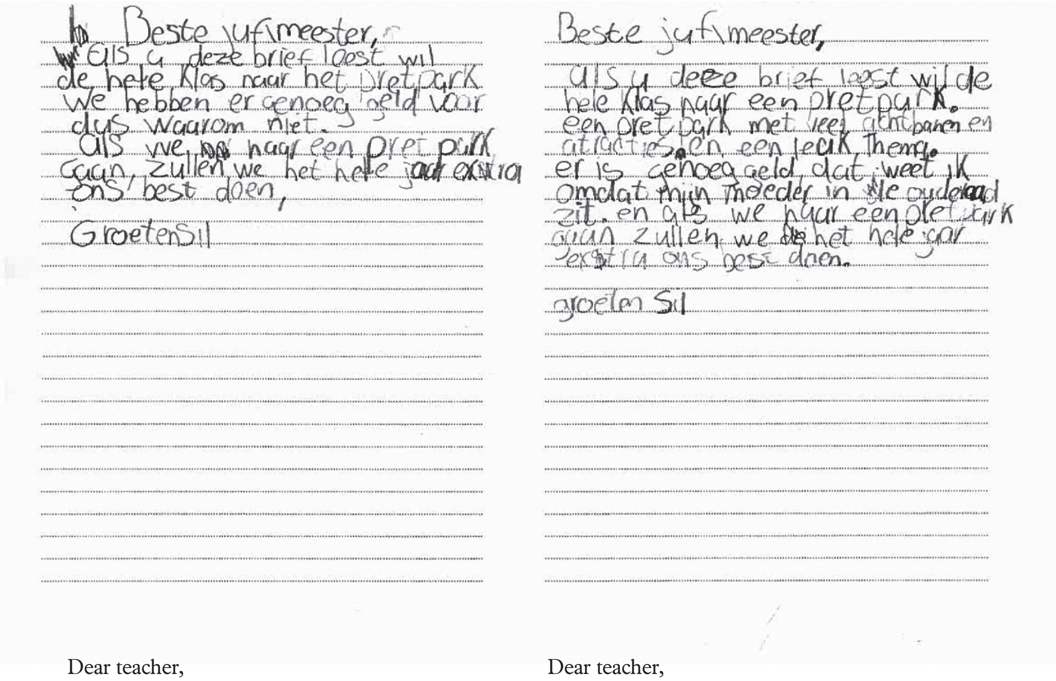
liefe groet
van
Ruben

Dear teacher

we write this letter because the schoolyard does not really have nice playing equpment. We would like to have a place where you can play with your friends and we have a large preference for the swings. but we also understand that your school does not have a lot of money so that is why we do not really mind, but it is the case that there will be fewer children at you school because of that and that the children do not even enjoy the breaks the most. so would you please look for a solution

deer greeting
by Ruben

Figure 10.3a An example of prewriting (left pane) and the final text (right pane) by the same student, written before the intervention.



Beste juf\meester,
als u deze brief leest wil
de hele klas naar het pretpark
we hebben er genoeg geld voor
dus waarom niet.
als we naar een pret park
gaan, zullen we het hele jaar exstra
ons best doen,

Groeten Sil

Beste juf\meester,
als u deze brief leest wil de
hele klas naar een pretpark.
een pretpark met veel glijbanen en
atracties. en een leuk thema.
er is genoeg geld dat weet ik
omdat mijn moeder in de ouderaad
zit. en als we naar een pretpark
gaan zullen we het hele jaar
exstra ons best doen.

groeten Sil

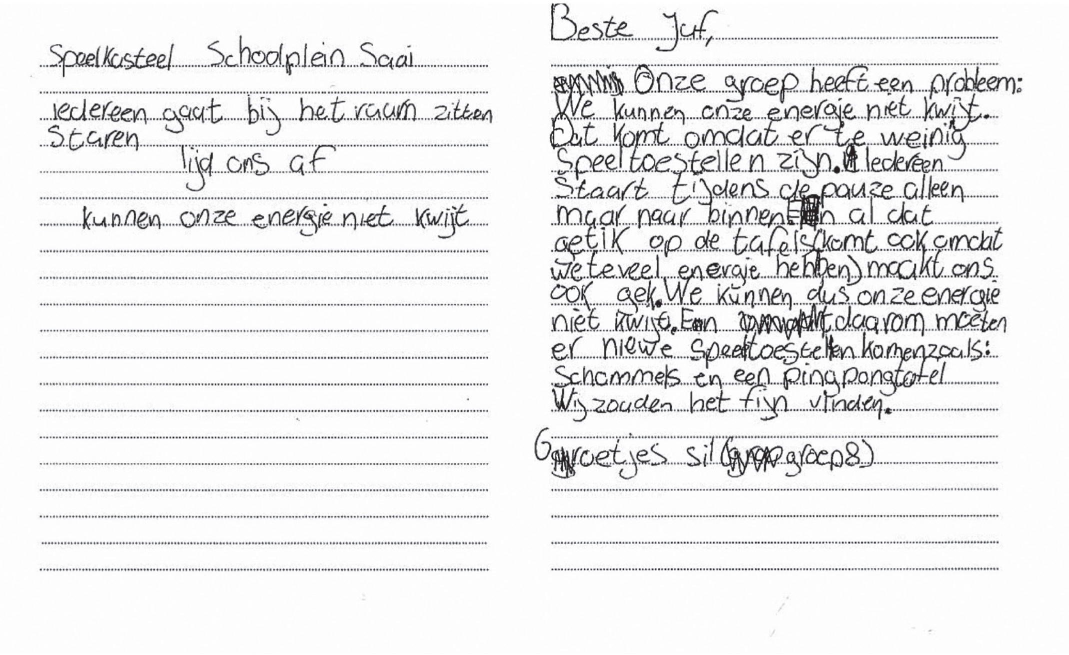
Dear teacher,
If you read this letter the whole class wants to go to the amusement park We have enough money so why not.
If we go to an amusement park, we will work exstra hard the whole year,

Regards Sil

Dear teacher,
If you read this letter the whole class wants to go to the amusement park. an amusement park with many slides and atractions. and a fun theme. I know there is enough money because my mother is in the parents' council.
and if we go to an amusement park, we will work exstra hard the whole year.

regards Sil

Figure 10.3b An example of prewriting (left pane) and the final text (right pane) by the same student, written after the intervention.



Speelkasteel Schoolplein Saai

Iedereen gaat bij het raam zitten
Staren
lijd ons af

kunnen onze energie niet kwijt

Beste Juf,

Onze groep heeft een probleem:
We kunnen onze energje niet kwijt.
Dat komt omdat er te weinig
Speeltoestellen zijn. Iedereen
Staart tijdens de pauze alleen
maar naar binnen. En al dat
getik op de tafels (komt ook omdat
we teveel energje hebben) maakt ons
ook gek. We kunnen dus onze energie
niet kwijt. En daarom moeten
er niewe speeltoestellen komen zoals:
Schommels en een pingpongtafel
Wij zouden het fijn vinden.

Groetjes sil (groep8)

playing castle playground boring

everyone is staring at the window

distrect us

cannot get rid off our energy

Dear Teacher,

Our class has a problem: We cannot get rid off our energy. That is because there is too little playing equipment. During the break everyone just stares inside And all the tapping on the tables (also because we have too much energy) drives us crazy too. So we cannot get rid off our energy. And that's why new playing equipment is needed like: swings and a pingpong table.
We would really like it.

Regards sil (grade 6)

This strategy divided the writing process into (1) the prewriting phase, in which students generate and subsequently organize ideas; (2) the writing phase, in which students translate their (organized) ideas into text; and (3) the postwriting phase, in which students reread, evaluate, and revise their texts.

In Chapter 6 and 7 we focused on how the strategy influences students' knowledge and their writing process, and how this in turn might lead to higher quality texts. In Chapter 6 we examined the effect of explicitly teaching a writing strategy on students' knowledge of writing, especially on their metacognitive knowledge. During a short intervention of three Tekster-lessons students were taught the writing strategy and were provided with specific instruction and opportunities to practice in how to apply the writing strategy to different writing tasks. Already after three lessons, students' knowledge about writing was enhanced dramatically, as reflected in the amount of writing advice that students gave in a letter to a fictitious peer. Overall, most of the advice in the letters pertained to punctuation, capitals, spelling, or grammar lower, which demonstrated that students' knowledge mainly concerned the lower order aspects of writing. However, the Tekster-lessons led to a significant increase in knowledge about the writing process and organizational aspects of writing in the intervention group. It should be noted that by using a letter of advice as an indication of writing knowledge, we measured writing knowledge indirectly, and measured primarily knowledge that is easy to verbalize. This does have its drawbacks: there were students who formulated advice that they clearly did not use themselves, e.g., stress the use of capitals while not using any in their own letter. Further, it may well be that students have and apply strategic knowledge without consciously paying attention to it. Nevertheless, the writing advice in the letters provided insight in the type of knowledge students in general possess. Further, our findings demonstrate that the more students knew about writing, specifically about the writing process, the better the quality of their texts was. In other words, explicitly teaching students a writing strategy enhanced students' metacognitive knowledge about how to write a text, which affected their writing performance significantly.

In Chapter 7 we primarily focused on the effectiveness of one important aspect of the writing strategy students were taught in Tekster, i.e., prewriting. In Tekster students were taught to generate ideas and organize these ideas before starting to write a full text. This form of prewriting specifically aimed to reduce students' cognitive overload during the writing process, which supposedly leads to better writing performance. The assumption is that if students have already planned what to write, they have all their attentional capacity left to focus on translating these ideas into text. Results from our study showed that, compared to a control group, students who received the intervention were not only more likely to use prewriting, but also used more effective prewriting strategies. The likelihood that students generated and organized their ideas multiplied at least 6 times after the intervention. In general, using a prewriting strategy led to texts of a higher quality. Generating and organizing ideas before writing was the most effective type of prewriting (ES = 0.68), in comparison to students who did not apply prewriting. Just listing ideas was slightly less effective (ES = 0.43). Writing a full draft was the least effective type of prewriting (ES = 0.23). Figure 10.3a and 10.3b show how the type of prewriting changed due to the intervention, and how this affected the quality of the final text.

Another key component that was implemented in Tekster is feedback. The meta-analysis in Chapter 2 showed that feedback is a highly effective instructional practice for learning to write. Whereas whole-classroom instructional practices are generally aimed at the average student, the purpose of feedback is to provide individualized instruction in order to promote learning for students with different proficiency levels. However, to provide effective feedback, teachers have to be able to adequately address students' individual needs. In Chapter 8 we investigated teachers' feedback practices in more detail. The results in this chapter showed that the type and amount of feedback depended on the preferred style of the teacher, rather than on the actual performance of the student. Teachers appeared to have one general preferred style of providing feedback, in which they primarily focused on either higher or lower concerns and gave their feedback in the same directive way; they hardly adjusted their feedback to students' individual needs. Based on the results of this study, it can be concluded that teachers need additional training to optimize their feedback to the needs of individual students, and hence, to provide the optimal possibility for weak as well as proficient students to improve their writing performance. As a consequence, feedback was a substantial part of the professional development program we developed to optimize the effectiveness of Tekster.

Intervention study 2: Teachers' implementation of Tekster

The findings of the intervention study reported in Chapter 5 showed that there were large differences between classes that could possibly be attributed to differences in teachers' ability to teach writing. Therefore, we conducted a second large-scale intervention study in which we added a professional development component to the writing intervention, which we described in Chapter 9. Comparable to the first intervention study in Chapter 5, teachers delivered the intervention program Tekster to their own students. In total, 68 teachers and 1365 students from 25 different schools from all over the Netherlands participated. In this study the duration of the program was extended to four months, with one lesson a week. Teachers received two training sessions, one prior to the start of the program and one during the course of the program. The training program component aimed at increasing teachers' pedagogical content knowledge and skills for teaching writing. To investigate whether the content of such a professional development training would be easily transferable among teachers we adopted a teachers-training-teachers approach in which half of the teachers were trained by experts, after which these teachers subsequently trained their colleagues. Comparable to the design of the first intervention study in Chapter 5, students' writing performance was assessed with a switching panel design with two groups and three measurement occasions in which students wrote three texts in different genres. The quality of each written text was rated by three raters using a benchmark rating procedure.

Results showed that the intervention program improved students' writing (ES = 0.55), and that the intervention was equally effective for both students in the teacher-trainer group and students in the teacher-trainee group. Already after four months of working with Tekster, students progressed more than one-and-a-half grade level. To measure the effectiveness of the additional training component that was included in the intervention, we used a mixed methods approach in which we triangulated information from teacher questionnaires, teacher interviews, and classroom observations. With this approach we aimed to get more information on how teachers' experienced the professional development activities, but also on how teachers' attitudes, self-efficacy for teaching writing, and their classroom practice were affected by the intervention. Results showed that the provided training was effective for teacher-trainers as well as teacher-trainees. Both trainers and trainees became more positive and felt more efficacious about teaching writing, and they were highly satisfied with the overall program and the professional development activities. Moreover, trainers and trainees were able to change their focus as well as their mode of instruction to the same extent, indicating that professional development is easily transferable between teachers. All in all, this study provides valuable clues on how teachers can effectively implement Tekster in their classroom and, thus, how the gap between research and classroom practice can be bridged to improve writing education.

DISCUSSION AND IMPLICATIONS FOR FURTHER RESEARCH

Explanations for students' improvements in writing

The results in the present dissertation show in a series of experiments that the writing program Tekster was highly effective in improving students' writing performance. The program aims to address the complexity of both writing and learning to write. Writing is cognitively challenging, especially for inexperienced writers, as several resource-demanding cognitive activities have to be performed simultaneously (Fayol, 1999). Learning to write poses the double challenge of writing and learning from this how to write at the same time. Unlike oral language skills, writing is not a skill you simply learn by doing (Rijlaarsdam et al., 2011). To become a proficient writer, a student needs knowledge about the writing process and written products, as well as ample opportunities for guided practice. It is essential that teachers, besides good teaching material, also possess the requisite skills and knowledge to support their students adequately during the writing lessons. Therefore Tekster aimed to improve the current situation in writing education in two ways: (1) by developing teaching material based on practices that have proven to be effective in previous research, and (2) by optimizing teachers' classroom practice in teaching writing through enhancing their skills and knowledge by a professional development program.

We decided to combine different approaches into one comprehensive program as in a typical classroom the instructional needs of students differ. These different instructional needs can be addressed by variation in instructional approach. In short, the instructional focus of Tekster includes three key principles: strategy-instruction, text-structure instruction, and self-regulation including goal-setting. According to our meta-analysis (see Chapter 2) these are effective practices to improve elementary students' writing performance. The instructional mode also includes three key principles: explicit instruction, observational learning, and (guided) practice with gradual release of responsibility from teacher to student. Together, these design principles for the focus and mode of instruction are translated into different learning and teaching activities, which form the blueprint of the teaching program Tekster (see Table 9.2). For example, in the Tekster-lessons students learn a strategy that subdivides the writing process into separate steps by observing a teacher modeling the strategy, listening actively to explicit instructions of the teacher on the purpose and benefits of using a strategy and applying the steps of the strategy to writing tasks themselves, while teachers provide help when needed through scaffolding and process feedback. Teachers were supported in the implementation of these instructional practices through detailed lesson plans in the teacher manual, a dvd with exemplary videos, and additional training sessions.

By combining several didactical approaches focusing on both the student and teacher, we believe that the program as a whole might be more effective than the sum of its parts. The results from the intervention studies did not show an aptitude-treatment interaction, indicating that the program as a whole was effective for the whole range of students: poor, average, as well as proficient writers. To gain more insight into how writing education can further be optimized, it is important to not only analyze the effectiveness of the program as a whole, but also what the effectiveness is of specific components of the program. Below we will try to explain in more detail the effectiveness of the program by focusing on components of the lesson material and the role of teachers respectively.

Effective components of the teaching material

Pre- and postwriting strategies. First, our findings show that strategy-focused instruction made students more consciously aware of their own writing process (Chapter 6) and changed their writing process (Chapter 7). Specifically, it was demonstrated that an increase in students' knowledge about the process as well as in their use of prewriting strategies led to texts of higher quality. These findings are in line with previous research that indicated that beginning writers predominantly employ a knowledge-telling approach to writing (Bereiter & Scardamalia, 1987): They immediately start writing and generate content on the go. As they have to think about all aspects simultaneously during writing (what to write, how to organize, how to spell words and use the right grammar, how to apply genre conventions, etc.), there is often a cognitive overload that hampers writing a text of good quality. By approaching the writing task in a more structured way: first think, and then act, students can subdivide the writing process in smaller components, reducing the cognitive overload during writing. The results in Chapter 7 further indicated that when students did not only generate ideas, but also organize their ideas before starting to write a text, they wrote even better developed and organized texts. This strategy of generating - organizing - writing was exactly what students learned as a prewriting strategy in Tekster. The findings in the two large-scale intervention studies also seem to support the notion that the writing strategy improved students' self-regulation skills. Although students were not explicitly instructed to use the writing strategy in the tasks they had to write at the consecutive measurement occasions, they nevertheless transferred the newly acquired prewriting strategy from the highly pre-structured lessons to these less structured writing assignments. Further, we have some indication of transfer to other subjects, as one of the students mentioned also using the writing strategy to write a history paper. In interviews teachers confirmed that students experienced much support by approaching the writing task step by step. They mentioned that it was much easier for students to start with the writing task, and that they wrote more because they were able to generate more ideas. Together, these results are a strong indication that the prewriting strategy is one of the most effective ingredients in Tekster.

Although the results with regards to prewriting are quite straightforward, the effects regarding postwriting are less clear. In Tekster, students are taught that their text is not finished when they have written the first full version, but that they subsequently have to reread, evaluate and revise their text. This postwriting strategy aimed to further improve the text after a first draft is written. However, in practice, students tended to skip the revision phase. It is not clear why they struggled to adopt this postwriting strategy: Was this because they lack the skills to critically evaluate their text or were they not motivated to revise a text that is already 'finished'? Preliminary results on students' revision behavior indicate that the likelihood that students (successfully) revised their text depended on the content and form of teachers' feedback (Bogaerds-Hazenberg, 2015). It was shown that students, on average, used only one third of the feedback for revision, and although more feedback was related to more revisions, this was not related to the quality of the revised text. Feedback that focused specifically on higher-order aspects of the text and provided explicit guidance for students for improvement supported students in successfully revising their text. These results suggest that feedback of the teacher is an important factor in optimizing students' revision behavior. However, from the findings in Chapter 8 it can be concluded that teachers have their own preferred style of providing feedback, regardless of students' individual needs. Together, these findings call for further research in which the effect of feedback on students' revision behavior should be investigated. Another way to make students more aware of the importance of revision, and to enhance their revision skills, is by using video clips in which peers model the revision process (cf. Van Steendam et al., 2010) or by observing how readers respond to their text (see Chapter 2 and 5). Further research needs to examine whether these interventions increase the likelihood that students engage in post-writing activities, and whether this leads to further improvement of their writing performance.

One limitation of our research is that we cannot be definitely sure that improvements in text quality are actually due to changes in the writing process, as students' progress was primarily assessed on the basis of the quality of their written products. Even in Chapter 7, in which we demonstrated that students adapted their writing process by using more prewriting, in a more effective way, this was determined by analyzing the written products from this particular stage of the writing process, i.e., the drafting space on the writing assignment. This does, however, not provide any information on the underlying processes, such as the decisions the writer makes on what to include in the text and how. Further, analyzing the product provides no insight on how the writing process developed over time. Van den Bergh & Rijlaarsdam (1996) demonstrated that text quality depends on the stage in the writing process in which a particular pre- or post-writing activity is employed. For instance, they established that planning activities have a positive effect on text quality in the beginning of the writing process, but this relation becomes negative at the end of the writing process. This pattern has also been established for younger writers (Van der Hoeven, 1997). Moreover, as the writing process of experienced writers is recursive rather than linear (Flower & Hayes, 1980), it could also be possible that more ideas are generated as the text is developing and not necessarily only during the prewriting phase. For further research, we therefore recommend to examine students' writing process in more depth in order to understand how they adopt and implement a new writing strategy. Information regarding the writing process can be obtained, for instance, by methods such as think-aloud protocols or triple-task paradigms in which the writer has to verbalize every writing activity (Olive, Kellogg, & Piolat, 2002) or by analyzing keystrokes (Leijten & Van Waes, 2013).

Instruction in criteria for good writing. The effectiveness of Tekster might also be explained by explicit instruction aimed at improving students' knowledge about what constitutes good writing. For instance, in Chapter 6 was shown that the intervention made students more aware of the importance of the organization of texts, which is regarded as a higher order aspect of writing. This improvement in students' knowledge about the written product might have been induced by different components in the writing program. For instance, in every Tekster-lesson students received explicit instruction in text structures, related to the genre of the text, or they discussed model texts illustrating texts of poor as well as good quality. Furthermore, at the end of each lesson they discussed their texts with peers, through which they experienced how a reader responds to their written text. This provided valuable clues about essential text characteristics. Hence, although we have established that the intervention increased students' knowledge of the written product, we cannot be sure whether this is caused by one learning activity in particular or by a combination of activities. Additional research is needed to examine this further.

Focus on the communicative aspect of writing. Another effective component in Tekster might be the explicit attention for the communicative function of writing. Each Tekster lesson starts with an introduction on the communicative goal of the writing task, and ends with an evaluation whether this goal is attained. Because most of the writing lessons have functional writing assignments with a clear communicative goal and intended audience, students can easily assess whether the text functions as intended. For example, students learn how to write instructions for a game or how to write a recipe. After the text is written, students are instructed to play the game or use the recipe of one of their peers, by which they experience how a reader responds to their text, which makes the communicative aspect of writing 'visible'. Although we did not specifically test the effect of communicative goals of writing and the use of functional writing assignments, we learned from interviews with teachers that students seem to be more motivated to write because they were more aware of the importance of good writing skills for communicative purposes.

Observational learning. In the way writing education in school is currently organized, learning to write and task execution are inextricably linked, which poses a double challenge for students: They have to produce a text and learn from this activity how to write at the same time. To separate learning from task performance, we included observational learning as one of the instructional practices in Tekster. With observational learning students observe others performing complex and unfamiliar tasks and gain insight into how a writing task can be approached. We applied observational learning in the Tekster-lessons in various ways: through teacher as well as peer modeling, and during two moments in the writing process, i.e., before and after writing. Observational learning was applied before writing to prepare students for the task at hand, e.g., by modeling how to generate ideas, organize them and translate them into text. After writing observational learning was applied to make students aware how a reader responds to their text, thus providing them with clues for revision. Although our program demonstrates how various forms of observational learning can be operationalized in a teaching program for writing, we did not investigate separately the effect of observational learning or the effects of different types of observational learning. However, interviews with teachers indicated that observing a model supported students to overcome difficulties during the writing process, such as progressing from one step of the strategy to the next. Further, the interviews suggested that by observing a reader reacting to their text, students became more aware of the communicative effectiveness of their text. As a consequence, students might be better prepared for the writing task they have to execute.

To gain more insight into the effectiveness of these forms of observational learning, further research is needed that specifically investigates which type of observational learning is effective for what type of students. For instance, prior research suggests that good students learn more from good models, whereas weak students learn more from weak models (Braaksma, 2002). Previous research also found that peer modeling is more effective in enhancing students' self-efficacy (Raedts et al., 2007; Schunk, 1987). To examine this into more depth, studies are needed in which different types of observational learning are compared with respect to their impact on the writing process as well as to the written product. The designated way to investigate this would be by applying a mixed methods approach, in which students' writing process as well as the quality of modeling are closely scrutinized.

The crucial role of the teacher

Effective writing education does not only depend on good teaching material, it also largely relies on the teacher. Results from the first intervention study showed large differences between classes in students' writing progress. Hence, some teachers were more effective in improving their students' writing performance than others, which is hardly surprising. In order to work effectively with Tekster, teachers have to be able to model the writing process, provide explicit instruction on how to write a good quality text, as well as facilitate guided practice during students' writing activities, monitor students' progress over time, and provide differentiated feedback. To optimize their writing instruction teachers need a combination of content and pedagogical knowledge: they need to know the dynamics of the writing process and what constitutes good writing, as well as know how they can transfer their knowledge to their students and scaffold students' learning in the most effective way.

To promote teachers' skills and knowledge for teaching writing and reduce differences between teachers, we included a professional development program in the second intervention study. The results from this study are promising, as teachers seemed to improve the way they teach writing. In general, questionnaires revealed that teachers were generally more positive about teaching writing after the intervention than before. Moreover, the intervention made them feel more efficacious to successfully teach writing and offer support to inexperienced writers. The interviews also revealed that the vast majority of teachers felt better equipped to teach writing. Especially their evaluations regarding the Tekster writing strategy were positive: they stated that the clear steps explicated in each writing lesson helped them to guide their students through the writing process. They also indicated that they experienced much support from the lesson plan in the teacher manual, which offered suggestions for explicit instruction and scaffolding. That teachers improved their writing education was also confirmed by observations of teachers' classroom practice. Before working with Tekster, teachers' instructional focus was mostly directed to the product. However, as a result of the writing program their focus had shifted to the writing process. For instance, they started to model (parts of) the writing process and used the writing strategy to teach students how to optimize their writing process. Moreover, they offered support to students throughout the whole lesson, and provided feedback on both the writing process and the written product. Taken together, these results show that, in general, teachers effectively changed their focus and mode of writing instruction.

Because we implemented the professional development program by a teachers-training-teachers approach, we were also able to examine whether teachers could effectively transfer newly acquired skills and knowledge to their colleagues. This transfer was successful, as we could not assess any differences between trainers and trainees in how they implemented the intervention, nor in the effect of the intervention on their students' writing. This was quite remarkable, however, as questionnaires and interviews revealed that the collegial training sessions primarily addressed the writing strategy and organizational aspects of implementing the writing program in the classroom, instead of the other key principles underlying the focus and mode of the teaching program. It is promising that teacher-trainees were still able to implement the intervention effectively, even with the 'light' version of the professional development program. It could be that teacher-trainers might have needed more time to familiarize themselves with the program before being able to transfer knowledge and skills regarding the instructional mode to their colleagues. This hypothesis should be the subject of further investigation in the future.

Even after following a professional development program, substantial differences between classes remained. This might be the result of differences between classes at pretest level due to differences in the composition of the class population; however, it might also be that, even though teachers changed their instruction while teaching writing, the quality of their instruction still differed considerably. For example, in Tekster, teachers are instructed to scaffold their students towards self-regulated writing. Initially, teachers provide explicit instruction and model (parts of) the writing process. Then, through scaffolding and feedback the responsibility is gradually released from teacher to student. From interviews and observations we learned that, to a large extent, teachers were able to change the organization of their lessons in such a way that self-regulated learning was promoted. However, teachers indicated that they sometimes struggled with modeling the writing process and with differentiating between less- and high-proficient writers. This suggests that teachers need more time and practice to master these skills.

Despite all our teacher measures, both quantitative and qualitative ones, we only registered whether teachers applied the key components of the program in their writing instruction; we did not measure the quality of the changes in teachers' classroom practice. Measuring the quality of changes in teachers' classroom practice asks for more detailed analyses which was hardly possible on such a large scale. The present research showed that, considering the large differences between teachers, a sufficient number of teachers should be included in order to get a good impression of the results of a professional development program in a regular classroom situation. For further research it is necessary to investigate how professional development of teachers should be designed and implemented in order to change teachers' focus as well as their mode of instruction in the most effective way. Regarding the professional development program it should be noted that we only examined one particular form of professional development. Hence, our results cannot be generalized to professional development in general.

Longitudinal development of students' writing skills

This dissertation demonstrates that Tekster improves students' writing performance significantly, and that these improvements are maintained even two months after following the writing program. In both intervention studies there was an effect of the number of lessons that were completed: the more lessons were completed, the more progress students made. Based on these findings, we might assume that when the program is used a full academic year, students' writing performance will improve to an even larger degree. However, more research is needed to establish this as the duration of the intervention in the present research was only two (Chapter 5) and four months (Chapter 9) respectively.

Besides this, the current results do not provide any insight in how the program might affect the longitudinal development of writing skills over grades as the intervention was only tested in a cohort study. It is of importance, however, to understand whether students' progress continues after one year of working with the program, as a good and effective writing program does not improve performance at one specific point in time (i.e., weeks or months), but follows a learning trajectory over multiple years aiming at continuous progress in students. This calls for further (longitudinal) research in which students who are using the writing program are followed across the grades. It would be particularly interesting to examine how much progress can be achieved, and how this progress develops, when students work with Tekster during multiple, consecutive years.

Until now, it has been difficult to measure progress in writing performance, as the measurement of writing was often very task-specific, which made comparison of writing performance across tasks and over time problematic. However, as demonstrated in Chapter 4, when the quality of students' writing is measured with a benchmark rating scale, this scale can be used to measure writing quality for different tasks. This would allow us to compare the quality ratings of multiple texts. Moreover, because a benchmark scale represents different performance levels that can be accomplished, it would be possible to monitor students' progression over time.

METHODOLOGICAL CONSIDERATIONS

Assessment of students' writing performance

The research presented in this dissertation not only aimed to develop and test the effectiveness of a new program for teaching writing, but also aimed to improve the assessment of writing. Specifically, we established that more information about students' writing performance than only one written text is necessary in order to draw inferences about students' underlying writing proficiency, and we developed a benchmark rating procedure that lead to more reliable and valid ratings of text quality than what is presently achieved. Increasing our understanding about the assessment of writing is essential for improving writing education for three reasons. First, teachers need to know whether their students perform at a sufficient level. Second, teachers can adapt their lessons to better suit the specific needs of students, either through general classroom instruction, or through individual feedback. Third, reliable and valid assessment of writing performance is necessary to monitor students' writing progress over time.

As in every study including writing assessments, we had to develop a rating procedure in which multiple raters would evaluate the quality of a large number of students' written texts, without exceeding raters' physical or psychological boundaries. The two large-scale intervention studies presented in the current dissertation were especially challenging, as they included a total of 2766 participating students who each wrote nine texts, varying in topic and genre. This resulted in a total of almost 25000 texts of which the quality needed to be rated by at least three raters. Of course, it was not possible to administer all texts to one jury of two or three raters. Even with a benchmark rating procedure, in which raters apparently only need approximately one minute to score the quality of each text, it would take too much time, which would possibly affect the reliability of their scores. Therefore, we decided to use overlapping rater teams (Van den Bergh & Eiting, 1989), in which each rater received three subsamples. In the design of overlapping rater teams writing products are randomly split into subsamples, equaling the number of raters. Each subsample is assigned to multiple raters (i.e., in this study to three raters) according to a prefixed design in which each rater is directly or indirectly linked to each of the other raters. Table 10.1 provides an example of such an overlapping design with five raters and five subsamples. There are various advantages related to the overlapping rater design. First, it decreases the total rating time for each individual rater. Second, the ratings are far more reliable and generalizable, because the scores do not depend on only one jury of raters, but are based on multiple juries (47 juries in study 1, and 18 juries in study 2). Third, the covariance matrix between raters allows for an estimation of the overall reliability, as well as the reliability per in-

Table 10.1 Design of overlapping rater teams with five raters and five subsamples

	Subsample				
	1	2	3	4	5
Rater 1	x	x	x		
Rater 2		x	x	x	
Rater 3			x	x	x
Rater 4	x			x	x
Rater 5	x	x			x

dividual rater. This estimation is based on the assumption that when subsamples are combined into one sample, none of the raters will know which text belongs to which subsample, rendering equal rater reliabilities across subsamples.

Appendix L presents an overview of the reliabilities of individual raters and juries of three raters of the two intervention studies presented in Chapter 5 and 9. In both studies it was shown that, while generalizing over all raters, who received only a short amount of training, the reliability of the benchmark rating procedure was quite satisfactory for individual raters (respectively $\varrho = .73$ and $\varrho = .71$) and high for juries of three raters (respectively $\varrho = .89$ and $\varrho = .88$). This confirms the findings from Chapter 4, which show that a benchmark rating procedure is a reliable and valid way to evaluate text quality.

There are, however, still some caveats associated with our assessment of students' writing performance. For instance, although students wrote three texts in different genres at each measurement occasion, this does not approximate the number of writing tasks that was recommended in Chapter 3. In this chapter it was demonstrated that, in order to generalize to writing proficiency, students have to write three different texts in each of four genres, equaling a total of 12 texts. In a regular educational setting, however, it is not feasible to include such a large number of writing tasks at one measurement occasion. In some schools, it would even outnumber the total number of writing tasks that students normally receive during the course of a school year (cf. Pullens, 2012). Therefore, we decided to limit the number of writing tasks in the assessment of writing to three. We purposely chose to include three writing tasks from different genres in order to maximize the task variance at each occasion, which provided us with a more generalized view on students' writing performance than when students would have written a single text, or multiple texts, in one genre. However, this is at the cost of a larger error component, which impedes possibilities to establish differences between conditions.

Further, although students' writing performance improved significantly, it is still unclear whether their performance meets the standards that are set by the Dutch Ministry of Education (Expert Group Learning Trajectories, 2009). In order to accomplish this, it is necessary that the benchmark rating scales do not only include different performance levels, but also that these levels match standardized levels that indicate the norms for sufficient and desired writing performance at the end of elementary education. Such a norm-procedure should be carried out by a team of experts, and is supposed to be of additional value for teachers, educational decision makers as well as researchers.

Generalizability and robustness of the results

Recently, social scientists highlighted the importance of replication (Open Science Collaboration, 2015). We tested the effectiveness in two large-scale intervention studies in two subsequent years among different samples of students, teachers and schools. In the first intervention study we implemented a switching replication design, which allowed us to test whether the effects of the intervention were reproducible. Taken together, the intervention was replicated in different groups, which yields important information about the generalizability and robustness of the findings.

The effect sizes in the present intervention studies (respectively 0.40 in Chapter 5 and 0.55 in Chapter 9 when teachers delivered the whole program of 16 lessons) are notably smaller in comparison to similar strategy-focused intervention studies aimed at grade 4 to 6 in a general educational setting (cf. Graham et al., 2012; Koster et al., 2015, average ES 1.02 and 0.96 respectively). In contrast to most previous writing intervention studies, we tested the intervention among a large number of students, teachers, as well as schools. In total, 2766 students and 144 teachers participated, from 125 classes and 52 schools. The schools varied with regards to their identity: there were regular public schools, as well as schools with a religious denomination (e.g., Catholic, Protestant, Reformed, or Islamic), and innovative schools (e.g., Montessori or Dalton). Because of this large number of students and teachers, as well as the variety of schools, we are able to generalize our conclusions across students, as well as across teachers and schools.

The vast majority of previous intervention studies did not incorporate such large numbers of schools, classes and students. Moreover, in previous studies students' performance is often assessed with only one writing task. Our results show that, if we would neglect the variance components related to tasks, the effect sizes of the present intervention studies would increase to 0.58 in intervention study 1 (Chapter 5) and 0.80 in intervention study 2 (Chapter 9). These effect sizes are comparable to the effect size of the intervention described in Chapter 7. This effect size, based on one task, was 0.72. The effect sizes of the intervention studies would increase even more if we also neglect the variance related to classes, i.e., 0.80 for intervention study 1 (Chapter 5) and 1.03 for intervention study 2 (Chapter 9). These effect sizes approximate the effect sizes reported in other intervention studies (cf. meta-analyses of Graham et al, 2012; Koster et al., 2015). Hence, although it might seem straightforward to directly compare effect sizes, the effect size can fluctuate depending on the sources of variance taken into account. The more variance components (e.g., schools, classes, students, tasks) are included in the analyses, the lower (but more realistic) the effect size will be.

In our intervention studies teachers delivered the intervention themselves, in their own classroom and with their own students. This is not regular practice in writing intervention research, as in most instances the intervention is not delivered by the teacher, but either by the researcher or a trained research assistant (cf. Chapter 9). Although this guarantees that the intervention is implemented as intended, one can hardly expect any maintenance effects of the intervention, as the teachers themselves do not change the way they teach writing. More importantly, when an intervention is delivered by a researcher, the generalizability of the results can only be limited to some controlled situation. Ecological validity can be increased by implementing the intervention in a naturalistic setting, as is the case in the present research, which makes it possible to generalize the findings to regular educational practice.

Intervention fidelity

As mentioned previously, the intervention was delivered by teachers in their own classrooms, in order to increase the ecological validity of the studies. Although this may facilitate bridging the gap between research and practice, there is a downside to this approach: it may be at the cost of the study's internal validity. If teachers deliver an intervention themselves, they often adapt it to their own classroom practice instead of exactly following lesson plans as prescribed. This could create a discrepancy between what researchers intended teachers to do and what they actually do. We have tried to minimize this discrepancy by involving teachers in every stage of the intervention study. Teachers were involved in the development of the lessons and we pretested a series of lessons in a pilot with a small sample of teachers, in order to ensure that the program was easy to use and that instructions were clear. Further, before teachers started with the program, they were instructed about the key components of the program: they were trained in how to include them in their writing lessons, and they received a detailed lesson plan for each lesson.

We also included fidelity measures in order to monitor whether teachers delivered the program as intended: we checked all student workbooks on the amount of lessons completed, and analyzed teachers' logbooks in which they had to register for each lesson their experience with the lesson and whether they were able to adhere to the lesson plans. Next to these quantitative measures, we also collected qualitative measures by observing teachers' classroom practice in a subsample of the classrooms and by interviewing teachers after the intervention. Although the fidelity measures revealed that teachers were able to execute the key components of the program, our measures did not shed light on the quality of the actual implementation in classroom practice. For instance, it might be that, although teachers modeled in each lesson, the quality of their modeling was rather low. We did not account for the quality of teachers' instructions in the observational checklists. Moreover, due to the large sample size, we did not observe in all classrooms and did not interview all teachers. For further research it is necessary to include fidelity measures that account for the quality of teachers' classroom practice as well. Such measures should give enough information to properly monitor and control teachers' classroom practice, but they also have to be applicable on a large scale, allowing for information on the quality of instruction of all participating teachers.

THE FUTURE OF TEKSTER: BRINGING EFFECTIVE WRITING INSTRUCTION INTO THE CLASSROOM

Although the present research and development project resulted in Tekster, an evidence-based program in writing, this, in itself, is not enough to make a difference in the writing education of tomorrow. In this paragraph we will present several recommendations to successfully implement Tekster in upper elementary grades.

First, teachers and school directors need to make a commitment to make writing a priority in their education and to invest time and resources to make their writing education more effective. An important prerequisite is that teachers acknowledge that writing, like reading and arithmetic, is one of the most fundamental skills that students need to master already in the early grades. Tekster has the potential to support teachers in boosting the writing skills of their students. As one of the participating teachers remarked: "the approach of Tekster is a recipe that always results in a delicious cake".

Further, it is of vital importance that tools for the assessment of writing are included in any program for the teaching of writing. And, of course, like the other fundamental skills such as reading or arithmetic, writing should be part of the national assessment program. Even though the development of a large-scale writing assessment is challenging, the benchmark rating procedure that is included in the writing program Tekster has shown promising results with regards to measuring writing in a reliable and valid way. Moreover, with regards to the procedure of assessing writing, it should become regular practice that students' writing proficiency is assessed with multiple writing tasks, rated by multiple raters. With only one task and one rater it is not possible to draw a solid conclusion about whether a student performs at a sufficient level. However, in the present educational system it is hard to organize juries of two or three raters: most of the time the classroom teacher evaluates the work of her own students. A benchmark rating scale can support the teacher in assessing and monitoring her students' progress. In contrast to a list of criteria, which only provides information on student's performance on a particular point in time and is often highly task-specific, a benchmark scale provides the teacher insight in the student's development over time and over tasks.

Last but not least, to provide future teachers with a solid foundation for effective writing instruction, the teaching of writing should be included in the pre-service curriculum. With regards to established teachers, our results have shown that even when only one teacher from one school receives professional development, through a teacher-training-teacher approach knowledge and skills can be transferred amongst colleagues within the school. Such an approach can be easily adopted in in-service training programs.

GENERAL CONCLUSION

This research project has demonstrated that writing education in the upper elementary grades can be improved through a combination of effective instructional practices. This combination resulted in an evidence-based program for teaching writing, Tekster, which proved to be easily implementable in daily classroom practice. Involvement of teachers in every stage of the developmental process ensured that the program fulfilled teachers' need for more support in teaching writing, based on a solid scientific foundation. Our findings prove that teaching students a writing strategy, combined with explicit text structure instruction, feedback and peer interaction is a very effective way to improve their writing performance. Together with more effective tools for the assessment of writing and a professional development program for teachers, Tekster lays an important foundation for upper elementary students' communicative skills.

REFERENCES

Alexander, P. A., Graham, S., & Harris, K. R. (1998). A perspective on strategy research: Progress and prospects. *Educational Psychology Review, 10*(2), 129-154.

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford, UK: Oxford University Press.

Ball, D. L., & Forzani, F. M. (2009). The work of teaching and the challenge for teacher education. *Journal of Teacher Education, 60*, 497-511.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bangert-Drowns, R.L., Hurley, M.M.,& Wilkinson, B. (2004). The effects of school-based Writing-to-Learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*, 29-58.

Barbeiro, L. F. (2011). What happens when I write? Pupils' writing about writing. *Reading and Writing, 24*(7), 813-834. doi:10.1007/s11145-010-9226-2

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86–107. doi:10.1016/j.asw.2007.07.001

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18,* 279-293. doi:10.1080/0969594X.2010.526585

Barkaoui, K. & Knouzi, I. (2012). Combining score and text analyses to examine task equivalence in writing assessments. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices* (Vol. 27, pp. 83-116). Leiden, The Netherlands: Brill.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and composition. *Journal of Reading Behavior, 16,* 297-306.

Beauvais, C., Olive, T., & Passerault, J. M. (2011). Why are some texts good and others not? Relationship between text quality and management of the writing processes. *Journal of Educational Psychology, 103*(2), 415–428. doi:10.1037/a0022545

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

Bereiter, C., Burtis, P. J., & Scardamalia, M. (1988). Cognitive operations in constructing main points in written composition. *Journal of Memory and Language, 27*, 261-278.

Berninger, V. W., Cartwright, A. C., Yates, C. M., Swanson, H. L., & Abbott, R. D. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades. *Reading and Writing, 6*(2), 161-196. doi:10.1007/BF01026911

Berninger, V. W., & Fuller, F. (1993). Gender differences in orthographic, verbal, and compositional fluency: Implications for assessing writing disabilities in primary grade children. *Journal of School Psychology, 30*(4), 363-382.

Berninger, V., Whitaker, D., Feng, Y., Swanson, H. L., & Abbott, R. D. (1996). Assessment of planning, translating, and revising in junior high writers. *Journal of School Psychology, 34*(1), 23-52. doi:10.1016/0022-4405(95)00024-0

Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing:*

An Interdisciplinary Journal, 4, 257-280. doi: 10.1007/BF01027151

Bhatia, V. K. (1993). *Analysing genre: Language use in professional settings.* London: Longman.

Biber, D., Nekrasova, T., & Horn, B. (2011). *The effectiveness of feedback for L1-English and L2-writing development: A meta-analysis* (Report No. RR-11-05). Princeton NJ: Educational Testing Service.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. doi:10.1080/0969595980050102

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

Blok, H. (1986). Essay rating by the comparison method. *Tijdschrift voor Onderwijs-research, 11*, 169–176.

Blok, H., & Hoeksma, J. B. (1984). Opstellen geschaald: De constructie van beoordelings-schalen voor vijf schrijfopdrachten [Scaling essays: The construction of rating scales for five writing tasks]. Amsterdam: Kohnstamm Institute.

Bodrova, E., & Leong, D. J. (1998). Scaffolding emergent writing in the zone of proximal development. *Literacy Teaching and Learning, 3*(2), 1-18.

Bogaerds-Hazenberg, S. T. M. (2015). *The influence of teacher feedback on revisions in children's writing* (Unpublished master's thesis). Utrecht University, the Netherlands.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2011). *Introduction to meta-analysis.* West Sussex: Wiley.

Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *Educational Researcher, 33*(8), 3-15.

Bourdin, B., & Fayol, M. (1994). Is written language production more difficult than oral language production? A working memory approach. *International Journal of Psychology, 29*(5), 591-620.

Bouwer, R., Béguin, A., Sanders. T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing, 32*, 83-100. doi:10.1177/0265532214542994

Bouwer, R., Koster, M., & Van den Bergh, H. (2016a). *Effects of a strategy-focused instructional program on the writing quality of upper elementary students.* Manuscript revised and resubmitted for publication.

Bouwer, R., Koster, M., & Van den Bergh, H. (2016b). *'Well done, but add a title!' Feedback practices of elementary teachers and the relationship with text quality.* Manuscript submitted for publication.

Bouwer, R., Koster, M., & Van den Bergh, H. (2016c). *Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner.* Manuscript revised and resubmitted for publication.

Braaksma, M. A. H. (2002). *Observational learning in argumentative writing* (Unpublished dissertation). University of Amsterdam.

Braaksma, M. A. H., Rijlaarsdam, G., Van den Bergh, H., & Van Hout-Wolters, B. H. A. M. (2010). Observational learning and its effect on the orchestration of writing processes. *Cognition and Instruction, 22*, 1-36. doi:10.1207/s1532690Xci2201_1

Braet, A., Moret, L., Schoonen, R., & Sjoer, E. (1993). Zo haal je een hoog cijfer voor je examenopstel: adviezen van en voor leerlingen: de perceptie van de doel stellingen van het opstelonderwijs in de bovenbouw van havo-vwo [That's how you get a good grade for your exam composition: perception of the aim of composition class in upper secondary education]. *Tijdschrift voor Taalbeheersing, 15*(3), 173-192.

Brandt, D. (1995). Accumulating literacy: Writing and learning to write in the twentieth century. *College English, 57*(6), 649-668.

Brannon, L., & Knoblauch, C. H. (1982). On students' rights to their own texts: A model of teacher response. *College Composition and Communication, 33*, 157–166.

Breland, H. M. (1983). *The direct assessment of writing skill: A measurement review* (Report No. 83-6). New York, NY: College Entrance Examination Board.

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

Brindle, M. (2013). *Examining relationships among teachers' preparation, efficacy, and writing practices* (Unpublished doctoral dissertation). Nashville, TN: Vanderbilt University.

Brodney, B., Reeves, C., & Kazelskis, R. (1999). Selected prewriting treatments: Effects on expository compositions written by fifth-grade students. *The Journal of Experimental Education, 68*(1), 5-20.

Broekkamp, H., & Van Hout-Wolters, B. (2007). The gap between educational research and practice: A literature review, symposium, and questionnaire. *Educational Research and Evaluation, 13*(3), 203-220.

Brunstein, J. C., & Glaser, C. (2011). Testing a path-analytic mediation model of how self-regulated writing strategies improve fourth graders' composition skills: A randomized controlled trial. *Journal of Educational Psychology, 103*, 922-938. doi:10.1037/a0024622

Bui, Y. N., Schumaker, J. B., & Deshler, D. D. (2006). The Effects of a Strategic Writing Program for Students with and without Learning Disabilities in Inclusive Fifth Grade Classes. *Learning Disabilities Research & Practice, 21*(4), 244-260.

Central Office for Statistics (2015, July 15). (Speciaal) basisonderwijs; culturele minderheden, (achterstands)leerlingen [(Special) primary education; cultural minority groups, (disadvantaged)students]. Retrieved from http://statline.cbs.nl/StatWeb/publication/?VW=T&DM=SLNL&PA=37846SOL&D1=0&D2=a&D3=a&D4=a&HD=090218-1354&HDR=T,G2,G1&STB=G3

Chai, C. (2006). Writing plan quality: Relevance to writing scores. *Assessing Writing, 11*(3), 198-223. doi:10.1016/j.asw.2007.01.001

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

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

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*, 65-81.

Clare, L., Valdés, R., & Patthey-Chavez, G. G. (2000). *Learning to write in urban elementary and middle schools: An investigation of teachers' written feedback on student compositions* (CSE Technical Report No. 526). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Coffman, W. E. (1966). On the validity of essay tests of achievement. *Journal of Educational Measurement, 3*(2), 151–156. doi:10.1111/j.1745-3984.1966.tb00872.x

Coffman, W. E. (1971). On the reliability of ratings of essay examinations in English. *Research in the Teaching of English, 5*(1), 24-36.

Connors, R. J., & Lunsford, A. A. (1993). Teachers' rhetorical comments on student papers. *College Composition and Communication, 44*, 200-223.

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

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Couzijn, M. J. (1995). *Observation of writing and reading activities: Effects on learning and transfer* (Unpublished dissertation). University of Amsterdam.

Couzijn, M. (1999). Learning to write by observation of writing and reading processes: Effects on learning and transfer. *Learning and Instruction, 9*, 109-142.

Couzijn, M., & Rijlaarsdam, G. (2004). Learning to write by reader observation and written feedback. In G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (Eds.), *Effective teaching and learning of writing. Current trends in research* (pp. 224-252). Amsterdam: Amsterdam University Press.

Covill, A. E. (1996). *Students' revision practices and attitudes in response to surface-related feedback as compared to content-related feedback on their writing* (Doctoral dissertation). Retrieved from Dissertation Abstracts International. (UMI No. 9716828)

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Los Angeles, CA: SAGE Publications.

Crismore, A. (1982). *Student perceptions of essential rules for successful academic compositions* (No. ED 221871). Retrieved from ERIC Document Reproduction Service: http://files.eric.ed.gov/fulltext/ED221871.pdf

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

Crossly, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research, 7*(3), 351-370. doi:10.17239/jowr-2016.07.03.02

Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education/Revue canadienne de l'éducation, 15*, 155-172. doi:10.2307/1495373

Crowhurst, M. (1991). Interrelationships between reading and writing persuasive discourse. *Research in the Teaching of English, 25*, 314-338.

Crowhurst, M., & Piche, G. L. (1979). Audience and mode of discourse effects on syntactic complexity in writing at two grade levels. *Research in the Teaching of English, 13*(2), 101–109.

De Glopper, K. (1988). *Schrijven beschreven. Inhoud, opbrengsten en achtergronden van het schrijfonderwijs in de eerste vier leerjaren van het voortgezet onderwijs* [Writing about writing. Content, results and backgrounds of the writing education in the first four years of secondary education] (Unpublished doctoral dissertation). The Hague: SVO.

De Smedt, F., Van Keer, H., & Merchie, E. (2015). Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing.* Advance online publication. doi:10.1007/s11145-015-9590-z

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

Department for Education. (2012). *What is the research evidence on writing?* (Research

Report No. DFE-RR238). Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/183399/DFE-RR238.pdf

Desimone, L. M. (2009). Improving Impact Studies of Teachers' Professional Development: Toward Better Conceptualizations and Measures. *Educational Researcher, 38*(3), 181-199. doi:10.3102/0013189X08331140

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). *Factors in judgments of writing ability* (Research Bulletin RB-61-15). Princeton, NJ: Educational Testing Service.

Dignath, C., & Büttner, G. (2008). Components of fostering self-regulated learning among students. A meta-analysis on intervention studies at primary and sec ondary school level. *Metacognition and Learning, 3*(3), 231-264. doi:10.1007/s11409-008-9029-x

Duijnhouwer, H., Prins, F. J., & Stokking, K. M. (2010). Progress feedback effects on students' writing mastery goal, self-efficacy beliefs, and performance. *Educational Research and Evaluation, 16*(1), 53–74. doi:10.1080/13803611003711393

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25,* 155-185. doi:10.1177/0265532207086780

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly, 9*, 270-292. doi:10.1080/15434303.2011.649381

Englert, C. S., Raphael, T. E., Anthony, L. M. A. H. M., & Stevens, D. D. (1991). Making strategies and self-talk visible: Writing instruction in regular and special education classrooms. *American Educational Research Journal, 28*(2), 337-372. doi:10.3102/00028312028002337

Englert, C. S., Raphael, T. E., Fear, K. L., & Anderson, L. M. (1988). Students' metacognitive knowledge about how to write informational texts. *Learning Disability Quarterly, 11*(1), 18-46.

Expert Group Learning Trajectories (2009). Referentiekader taal en rekenen: De referentieniveaus [Reference framework language and arithmetic: The referential levels]. Retrieved from the Ministry of Education, Culture and Science: http://www.taalenrekenen.nl/downloads/referentiekader-taal-en-rekenen-referentieniveaus.pdf

Fayol, M. (1999). From on-line management problems to strategies in written composition. In M. Torrance & G. Jeffery (Eds), *The cognitive demands of writing: Processing capacity and working memory effect in text production* (pp. 13-23). Amsterdam: Amsterdam University Press.

Feenstra, H. (2014). *Assessing writing ability in primary education. On the evaluation of text quality and text complexity* (Unpublished doctoral dissertation). University of Twente.

Feldt, L. S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests admnistered to the same sample. *Psychometrika, 45*, 99-105.

Ferrari, M., Bouffard, T., & Rainville, L. (1998). What makes a good writer? Differences in good and poor writers' self-regulation of writing. *Instructional Science, 26*(6), 473-488.

Ferretti, R. P., Lewis, W. E., & Andrews-Weckerly, S. (2009). Do goals affect the structure of students' writing strategies? *Journal of Educational Psychology, 101*, 577-589. doi:10.1037/a0014702

Ferretti, R. P., MacArthur, C. A., & Dowdy, N. C. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normally achieving peers. *Journal of Educational Psychology, 92*, 694-702. doi:10.1037//0022-0663.92.4.694

Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. *Assessing Writing, 19*, 6-23. doi:10.1016/j.asw.2013.09.004

Fidalgo, R,, Torrance, M., Rijlaarsdam, G., Van den Bergh, H., & Lourdes Álvarez, M. (2015). Strategy-focused writing instruction: Just observing and reflecting on a model benefits 6th grade students. *Contemporary Educational Psychology, 41*, 37-50. doi:10.1016/j.cedpsych.2014.11.004

Fife, J. M., & O'Neill, P. (2001). Moving beyond the written comment: Narrowing the gap between response practice and research. *College Composition and Communication, 53*, 300–321.

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research, 57*, 481-506.

Fitzgerald, J., & Markham, L. R. (1987). Teaching children about revision in writing. *Cognition and Instruction, 4*, 3-24. doi:10.1207/s1532690xci0401_1

Fitzgerald, J., & Teasley, A. B. (1986). Effects of instruction in narrative structure on children's writing. *Journal of Educational Psychology, 78*, 424-432. doi:10.1037/0022-0663.78.6.424

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

Flower, L., & 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, NJ: Erlbaum.

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

Follman, J. C. & Anderson, J. A. (1967). An investigation of the reliability of five procedures for grading English themes. *Research in the Teaching of English, 1*, 190-200.

Franssen, H. M. B., & Aarnoutse, A. (2003). Schrijfonderwijs in de praktijk. [Writing education in practice]. *Pedagogiek, 23*(3), 185-198.

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

Galbraith, D., Ford, S., Walker, G., & Ford, J. (2005). The contribution of different components of working memory to knowledge transformation during writing. *L1-Educational Studies in Language and Literature, 5*(2), 113-145. doi:10.1007/s10674-005-0119-2

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507–531.

Gillespie, A., Olinghouse, N. G., & Graham, S. (2013). Fifth-grade students' knowledge about writing process and writing genres. *The Elementary School Journal, 113*(4), 565-588.

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

Goldberg, A., Russell, M., & Cook, A. (2003). The effect of computers on student writing: A meta-analysis of studies from 1992 to 2002. *Journal of Technology,* *Learning, and Assessment, 2*(1), 1-52. Available from http://www.jtla.org.

Gordon, C. J., & Braun, C. (1986). Mental processes in reading and writing: A critical look at self-reports as supportive data. *Journal of Educational Research, 79*, 292-301. doi:10.1080/00220671.1986.10885694

Grabowski, J., Becker-Mrotzek, M., Knopp, M., Jost, J., & Weinzierl, C. (2014). Comparing and combining different approaches to the assessment of text quality. In D. Knorr, C. Heine, & J. Engberg (Eds.), *Methods in writing process research* (pp. 147-165). Frankfurt am Main: Peter Lang.

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds), *Handbook of writing research* (pp. 187-207). New York, NY: The Guilford Press.

Graham, S. (2013). *It all starts here. Fixing our national writing crisis from the foundation.* Colombus, OH: Saperstein Associates.

Graham, S., Harris, K., Fink, B., & MacArthur, C. (2001). Teacher efficacy in writing: A construct validation with primary grade teachers. *Scientific Studies of Reading, 5*(2), 177-202. doi:10.1207/S1532799Xssr0502_3

Graham, S., Harris, K., & Hebert, M. A. (2011). *Informing writing: The benefits of formative assessment. A Carnegie Corporation Time to Act report.* Washington, DC: Alliance for Excellent Education.

Graham, S., Harris, K., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology, 30*(2), 207-241. doi:10.1016/j.cedpsych.2004.08.001

Graham, S., MacArthur, C., & Schwartz, S. (1995). Effects of goal-setting and procedural facilitation on the revising behavior and writing performance of students with writing and learning problems. *Journal of Educational Psychology, 87*, 230-240. doi:10.1037/0022-0663.87.2.230

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104,* 879-896. doi:10.1037/a0029185

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445-476. doi:10.1037/0022-0663.99.3.445

Graham, S., & Sandmel, K. (2011). The process writing approach: A meta-analysis. *The Journal of Educational Research, 104*(6), 396-407.

Guskey, T. R. (1994). Results-oriented professional development: In search of an optimal mix of effective practices. *Journal of Staff Development, 15*, 42-42.

Hakstian, A.R. & Whalen, T.E. (1976). A K-sample significance test for independent alpha coefficients. *Psychometrika, 41*, 219-231.

Hamman, D., Berthelot, J., Saia, J., & Crowley, E. (2000). Teachers' coaching of learning and its relation to students' strategic learning. *Journal of Educational Psychology, 92*(2), 342-348.

Harris, K. R., & Graham, S. (1996). *Making the writing process work: Strategies for com position and self-regulation* (2nd ed.). Cambridge, MA: Brookline Books.

Harris, K. R., Graham, S, Brindle, M., & Sandmel, K. (2009). Metacognition and children's writing. In D. Hacker, J. Dunlosky, & A. Graesser (Eds.) *Handbook of metacognition in Education* (pp 131-153). New York: Routledge.

Harris, K. R., Graham, S., & Mason, L. H. (2006). Improving the writing, knowledge, and motivation of struggling young writers: Effects of self-regulated strategy development with and without peer support. *American Educational Research Journal, 43*(2), 295-340.

Harris, K. R., Graham, S., Mason, L. H., & Saddler, B. (2002). Developing self-regulated writers. *Theory into Practice, 41*, 110-115.

Harris, K. R., Lane, K. L., Graham, S., Driscoll, S. A., Sandmel, K., Brindle, M., & Schatschneider, C. (2012). Practice-based professional development for self-regulated strategies development in writing a randomized controlled study. *Journal of Teacher Education, 63*, 103–119. doi:10.1177/0022487111429005

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement.* London: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112. doi:10.3102/003465430298487

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. 29-57). Mahwah, NJ: Lawrence Erlbaum Associates.

Henkens, L. S. J. M. (2010). *Het onderwijs in het schrijven van teksten* (pp. 1–58) [Education in text writing]. Utrecht: Inspectorate of Education.

Hillocks, G. (1982). The interaction of instruction, teacher comment, and revision in teaching the composing process. *Research in the Teaching of English, 16*(3), 261–278.

Hillocks, G. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education, 93*, 133-170. doi:10.1086.443789

Hillocks, G. (1986). *Research on written composition: New directions for teaching.* Urbana, IL: ERIC Clearinghouse on Reading and Communication Skills.

Hintze, J. M., Volpe, R.J., & Shapiro, E.S. (2002). Best practices in systematic direct observation of student behavior. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 993-1006). Washington, DC: National Association of School Psychologists.

Hoetker, J. (1982). Essay examination topics and students' writing. *College Composition and Communication, 33*(4), 377–392. doi:10.2307/357949

Holliway, D. R., & McCutchen, D. (2004). Audience perspective in young writers' composing and revising. Reading as the reader. In G. Rijlaarsdam (Series Ed.), & L. Allal, P. Chanquoy, & P. Largy (Vol. Eds.), *Studies in Writing: Vol. 13. Revision: Cognitive and instructional processes* (pp. 105-121). Dordrecht: Kluwer Academic Publishers.

Huang, C. (2009). Magnitude of task-sampling variability in performance assessment: A meta-analysis. *Educational and Psychological Measurement, 69*(6), 887–912.

Huot, B. (1990a). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201–213. doi:10.2307/358160

Huot, B. (1990b). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research, 60*(2), 237–263.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning.* Logan, UT: Utah State University Press.

Hyland, F., & Hyland, K. (2001). Sugaring the pill: Praise and criticism in written feedback. *Journal of Second Language Writing, 10,* 185-212.

Inspectorate of Education (2012). *De staat van het onderwijs. Onderwijsverslag 2010/2011* [The state of education. Educational report 2010/2011]. Retrieved from http://www.onderwijsinspectie.nl/binaries/content/assets/Onderwijsverslagen/2012/onderwijsverslag_2010_2011_printversie.pdf

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2007). Toward a definition of mixed methods research. *Journal of Mixed Methods Research, 1*, 112-133. doi:10.1177/1558689806298224

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2,* 130-144. doi:10.1016/j.edurev.2007.05.002

Kellogg, R. T. (1988). Attentional overload and writing performance: effects of rough draft and outline strategies. *Journal of Experimental Psychology: Learning, Memory and Cognition, 14*, 355-365.

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

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*, 254–284.

Koopman, P., Ledoux, G., Karssen, M., Van der Meijden, A., & Petit, R. (2015). *Vervolgmeting 1 kengetallen passend onderwijs* [Sequal measure 1 key ratios inclusive education] (Rapport 936, No. 20667). Amsterdam: Kohnstamm Institute.

Kos, R., & Maslowski, C. (2001). Second graders' perceptions of what is important in writing. *The Elementary School Journal*, 567-584.

Koster, M., Bouwer, R., & Van den Bergh, H. (2014a). *VOS: werkboek en docentenhandleiding voor groep 6* [FOX: workbook and teacher manual for grade 4]. Utrecht: Utrecht University.

Koster, M., Bouwer, R., & Van den Bergh, H. (2014b). *DODO: werkboek en docentenhandleiding voor groep 7* [DODO: workbook and teacher manual for grade 5]. Utrecht: Utrecht University.

Koster, M., Bouwer, R., & Van den Bergh, H. (2014c). *EKSTER: werkboek en docentenhandleiding voor groep 8* [MAGPIE: workbook and teacher manual for grade 6]. Utrecht: Utrecht University.

Koster, M., Bouwer, R., & Van den Bergh, H. (2016a). *A letter of advice: The relationship between metacognitive knowledge and writing performance for elementary students.* Manuscript submitted for publication.

Koster, M., Bouwer, R., & Van den Bergh, H. (2016b). *Professional development of teachers in the implementation of a strategy-focused writing intervention program for elementary students.* Manuscript revised and resubmitted for publication.

Koster, M., Tribushinina, E., De Jong, P. F., & Van den Bergh, H. (2015). Teaching children to write: A meta-analysis of writing intervention research. *Journal of Writing Research, 7*(2), 249-274. doi:10.17239/jowr-2015.07.02.02

Kress, G. (1994). *Learning to write.* London: Routledge.

Krom, R., Van de Gein, J., Van der Hoeven, J., Van der Schoot, F., Verhelst, N., Veldhuijzen, N. & Hemker, B. (2004). *Balans van het schrijfonderwijs op de basisschool. Uitkomsten van de peilingen in 1999: halverwege en eind basisonderwijs en speciaal basisonderwijs* [Present state of writing competency in elementary education. Results of assessment in 1999: halfway and end of elementary and special education]. Arnhem: Cito.

Kuhlemeier, H., Van Til, A., Hemker, B., De Klijn, W., & Feenstra, H. (2013). *Balans van de schrijfvaardigheid in het basis- en speciaal basisonderwijs 2* [Present state of writing competency in elementary and special education 2] (PPON Report No. 53). Arnhem: Cito.

Langer, J. A., & Applebee, A. N. (1987). *How Writing Shapes Thinking: A Study of Teaching and Learning* (Research Report No. 22). Urbana, IL: National Council of Teachers of English.

Latham, G. P., & Locke, E. A. (1991). Self-regulation through goal-setting. *Organizational Behavior and Human Decision Processes, 50*, 212-247.

Lee, I. (2008). Ten mismatches between teachers' beliefs and written feedback practice. *ELT Journal, 63*, 13–22. doi:/10.1093/elt/ccn010

Lee, I. (2014). Feedback in writing: Issues and challenges. *Assessing Writing, 19*, 1-5. doi:10.1016/j.asw.2013.11.009

Lee, Y. W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing, 7*(4), 353–385.

Leijten, M., & L. Van Waes (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication, 30*, 358-392. doi:10.1177/0741088313491692

Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (in press). Comparative judgement as a promising alternative. In E. Cano & G. Ion (Eds.), *Innovative practices for higher education assessment and measurement.* Hershey, PA: IGI Global.

Lieberman, A., & Friedrich, L. (2007). Teachers, writers, leaders. *Educational Leadership, 65*(1), 42-47.

Limpo, T., Alves, R. A., & Fidalgo, R. (2014). Children's high-level writing skills: Development of planning and revising and their contribution to writing quality. *British Journal of Educational Psychology, 84*(2), 177-193. doi:10.1111/bjep.12020

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., . . . Busick, M. D. (2012). *Translating the statistical representation of the effects of education interventions into more readily interpretable forms* (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper and L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging.* Urbana, IL: National Council of Teachers of English.

Locke, E. A., Shaw, K. N., Saari, L. M., & Latham, G. P. (1981). Goal-setting and task performance: 1996-1980. *Psychological Bulletin, 90*, 125-152.

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

Luiselli, J. K., & Reed, D. D. (2011). Social validity. In *Encyclopedia of Child Behavior and Development* (pp. 1406-1406). Springer US.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246-276. doi:10.1191/0265532202lt230oa

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54-71.

Matsumura, L. C., Patthey-Chavez, G. G., Valdés, R. & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower-and higher-achieving urban schools. *The Elementary School Journal, 103*, 3–25.

McColly, W. (1970). What does educational research say about the judging of writing ability? *The Journal of Educational Research, 64*, 148-156.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25*(4), 431-444.

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

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research, 3*, 51-68.

McKeown, D., Fitzpatrick, E., & Sandmel, K. (2014). SRSD in practice: Creating a professional development experience for teachers to meet the writing needs of students with EBD. *Behavioral Disorders, 40*(1), 15-25. doi:10.17988/0198-7429-40.1.15

Meijerink, H. (2008). *Over de drempels met taal en rekenen* (pp. 1–56). Enschede: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Meuffels, B. (1994). *De verguisde beoordelaar: opstellen over opstelbeoordeling* [The derided rater: essays on essay rating]. Amsterdam, The Netherlands: Thesis Publishers.

Midgette, E., Haria, P., & MacArthur, C. (2008). The effects of content and audience awareness goals for revision on the persuasive essays of fifth- and eighth-grade students. *Reading and Writing, 21*, 131-151. doi:10.1007/s11145-007-9067-9

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63*, 81-96.

Ministry of Education, Culture and Science (2015). Basisonderwijs 2015-2016 [Primary Education 2015-2016]. Den Haag: Ministry of Education, Culture and Science.

Morphy, P., & Graham, S. (2012). Word processing programs and weaker writers/ readers: A meta-analysis of research findings. *Reading and Writing, 25*, 641-678.

Moss, P. A., Cole, N. S., & Khampalikit, C. (1982). A comparison of procedures to assess written language skills at grades 4, 7, and 10. *Journal of Educational Measurement, 19*, 37-47. doi:10.1111/j.1745-3984.1982.tb00113.x

Mullis, I. V. S. (1984). Scoring direct writing assessments: What are the alternatives? *Educational Measurement: Issues and Practice, 3*, 16-18.

Murphy, S (2000). A sociocultural perspective on teacher response: Is there a student in the room? *Assessing Writing, 7*, 79-90.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*, 386-422.

National Center for Education Statistics. (2012). *The nation's report card: Writing 2011.* Washington, DC: Institute of Education Sciences, U.S. Department of Education.

National Commission on Writing (2003). *The need for a writing revolution. The neglected "R".* New York: College Entrance Examination Board.

Neumann, A. (2012). DESI - Text Production. In M. Torrance, D. Alamargot, M. Castelló, F. Ganier, O. Kruse, A. Mangen. L. Tolchinsky, & L. van Waes (Eds.), *Learning to Write Effectively - Current trends in European Research* (pp. 193-197). Bingley: Emerald Group Publishing.

Northwest Regional Educational Library (2013). *About 6 + 1 Trait Writing.* Retrieved from http://www.educationnorthwest.og/resource/949.

Ofsted (2012). Moving English forward: *Action to raise standards in English.* Retrieved from: www.ofsted.gov.uk/resources/110118.

Olinghouse, N. G., & Graham, S. (2009). The relationship between the discourse knowledge and the writing performance of elementary-grade students. *Journal of Educational Psychology, 101*(1), 37. doi:10.1037/a0013248

Olinghouse, N. G., Graham, S., & Gillespie, A. (2015). The relationship of discourse and topic knowledge to fifth graders' writing performance. *Journal of Educational Psychology, 107*(2), 391. doi:10.1037/a0037549

Olinghouse, N. G., Santangelo, T., & Wilson, J. (2012). Examining the validity of single-occasion, single-genre, holistically scored writing assessments. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practices* (Vol. 27, pp. 55-82). Leiden, The Netherlands: Brill.

Olive, T., Kellogg, R. T., & Piolat, A. (2002). The triple task technique for studying the process of writing. In *Contemporary tools and techniques for studying writing* (pp. 31-59). Springer Netherlands.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. doi:10.1126/science.aac4716

O'Sullivan, J., & Pressley, M. (1984). Completeness of instruction and strategy transfer. *Journal of Experimental Child Psychology, 38*, 275-288.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*(3), 307-332. doi:10.3102/00346543062003307

Pajares, F., & Valiante, G. (2001). Gender differences in writing motivation and achievement of middle school students: A function of gender orientation? *Contemporary Educational Psychology, 26*(3), 366-381.

Paris, S. G., Lipson, M. Y., & Wixson, K. K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology, 8*(3), 293-316.

Parkes, J. (2001). The role of transfer in the variability of performance assessment scores. *Educational Assessment, 7*(2), 143–164.

Parr, J. M., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*, 68–85. doi:10.1016/j.asw.2010.05.004

Pearson, P. D., & Gallagher, G. (1983). The instruction of reading comprehension. *Contemporary Educational Psychology, 8*, 317-344.

Peterson, S. S., & McClay, J. (2010). Assessing and providing feedback for student writing in Canadian classrooms. *Assessing Writing, 15*, 86-99. doi:10.1016/j.asw.2010.05.003

Piolat, A., & Roussey, J. Y. (1996). Students' drafting strategies and text quality. *Learning and Instruction, 6*(2), 111-129.

Podsakoff, P. M., & Fahr, J. (1989). Effects of feedback sign and credibility on goal-setting and task performance. *Organizational Behavior and Human Decision Processes, 44*, 45-67.

Pollitt, A. (2012). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*, 281-300. doi:10.1080/0969594X.2012.665354

Pollman, E., Prenger, J., & De Glopper, K. (2012). Het beoordelen van leerlingteksten met behulp van een schaalmodel [Rating student' texts with a benchmark scale]. *Levende Talen Tijdschrift, 13*(3), 15-24.

Pullens, T. (2012). *Bij wijze van schrijven: Effecten van computerondersteund schrijven in het primair onderwijs.* [In a manner of writing: Effects of computer-supported writing in primary education] (Unpublished doctoral dissertation). Utrecht University.

Pullens, T., Den Ouden, H., Herrlitz, W., & Van den Bergh, H. (2012). Effecten van het computerprogramma TiO-schrijven op de schrijfvaardigheid van leerlingen in groep acht: een longitudinaal onderzoek. [Effects of intensive use of computer program TiO-schrijven on writing in primary education]. *Pedagogische Studieën, 89*, 104-119.

Quellmalz, E. S., Capell, F. J., & Chou, C. (1982). Effects of discourse and response mode on the measurement of writing competence. *Journal of Educational Measurement, 19*(4), 241–258. doi:10.1111/j.1745-3984.1982.tb00131.x

Raedts, M., Rijlaarsdam, G., Van Waes, L., & Daems, F. (2007). Observational learning through video-based models: Impact on students' accuracy of self-efficacy beliefs, task knowledge and writing performances. In G. Rijlaarsdam, P. Boscolo, & S. Hidi (Eds), *Studies in writing: Writing and motivation* (Vol. 19, pp. 219-238). Oxford: Elsevier.

Raphael, T. E., & Kirschner, B. M. (1985). *The effects of instruction in compare/contrast text structure on sixth-grade students' reading comprehension and writing products* (Research Series No. 161). East Lansing, MI: Michigan State University, Institute for Research on Teaching.

Raphael, T. E., Englert, C. S., & Kirschner, B. W. (1989). Students' metacognitive knowledge about writing. *Research in the Teaching of English*, 343-379.

Rau, P. S. (1996). How initial plans mediate the expansion and resolution of options in writing. *The Quarterly Journal of Experimental Psychology: Section A, 49*(3), 616-638. doi:10.1080/713755642

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1–32.

Reed, W. M., Burton, J. K., & Kelly, P. P. (1985). The effects of writing ability and mode of discourse on cognitive capacity engagement. *Research in the Teaching of English, 19*(3), 283–297.

Rietdijk, S., Van Weijen, D., Janssen, T., Van den Bergh, H., & Rijlaarsdam, G. (2015). *Teaching writing in primary education: Classroom practices, learning time, and teacher characteristics and their relationship*s. Manuscript submitted for publication.

Rietdijk, S., Van Weijen, D., Janssen, T., Van den Bergh, H., & Rijlaarsdam, G. (2016). *Writing strategy instruction: An intervention study.* Manuscript in preparation.

Rijlaarsdam, G. (2005). Observerend leren: Een kernactiviteit in taalvaardigheidsonderwijs [Observational learning: A core activity in language education.]. *Levende Talen Tijdschrift, 6*(4), 10-28.

Rijlaarsdam, G., & Couzijn, M. (2000). Writing and learning to write: A double challenge. In R. Simons, J. Van der Linden, & T. Duffy (Eds.), *New learning* (pp. 157-189). Dordrecht: Kluwer Academic Publishers.

Rijlaarsdam, G., Braaksma, M., Couzijn, M., Janssen, T., Raedts, M., Van Steendam, E., . . . Van den Bergh, H. (2008). Observation of peers in learning to write. practice and research. *Journal of Writing Research, 1*, 53-83.

Rijlaarsdam, G., Couzijn. M., Janssen, T., Braaksma, M., & Kieft, M. (2006). Writing experiment manuals in science education: The impact of writing, genre, and audience. *International Journal of Science Education, 28,* 203–233.

Rijlaarsdam, G., Janssen, T., Rietdijk, S., & Van Weijen, D. (in press). Reporting design principles for effective instruction of writing: Intervention as constructs. In R. Fidalgo, K. Harris, & M. Braaksma (Eds.), *Design principles for teaching effective writing: Theoretical and empirical grounded principles.* Leiden: Brill Publishers.

Rijlaarsdam, G., Van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., . . . Raedts, M. (2012). Writing. In K. Harrits et al. (Eds.), *APA educational psychology handbook: Application to learning and teaching* (Vol 3, pp. 189-227). Washington, DC: American Psychological Association. doi:10.1037/13275-009

Rogers, L., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology, 100*(4), 879-906.

Saddler, B., & Graham, S. (2007). The relationship between writing knowledge and writing performance among more and less skilled writers. *Reading & writing quarterly, 23*(3), 231-247. doi: 10.1080/10573560701277575

Saddler, B., Moran, S., Graham, S., & Harris, K. R. (2004). Preventing writing difficulties: The effects of planning strategy instruction on the writing performance of struggling writers. *Exceptionality, 12*(1), 3-17. doi:10.1207/s15327035ex1201_2

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 119-144.

Salahu-Din, D., Persky, H., & Miller, J. (2008). *The nation's report card: Writing 2007* (NCES 2008-468). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Scardamalia, M., & Paris, P. (1985). The function of explicit discourse knowledge in the development of text representations and composing strategies. *Cognition and Instruction, 2,* 1-39.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing, 22,* 1-30. doi:10.1191/0265532205lt295oa

Schoonen, R. (2012). The validity and generalizability of writing scores: The effect of rater, task and language. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (Vol. 27, pp. 1–22). Leiden: Brill. doi:10.1108/S1572-6304(2012)0000027004

Schoonen, R., & De Glopper, K. (1996). Writing performance and knowledge about writing. In G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (Eds.), *Theories, models, and methodology in writing research* (pp. 87-107). Amsterdam: Amsterdam University Press.

Schoonen, R., Van Gelderen, A., Stoel, R. D., Hulstijn, J., & De Glopper, K. (2011). Modeling the development of L1 and EFL writing proficiency of secondary school students. *Language Learning, 61*(1), 31-79. doi:10.1111/j.1467-9922 .2010.00590.x

Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing, 14*(2), 157–184.

Schriver, K. A. (1992). Teaching writers to anticipate reader's needs: A classroom pedagogy. *Written Communication, 9*(2), 179-208. doi:10.1177/0741088392009002001

Schunk, D. H. (1987). Peer models and children's behavioral change. Review of *Educational Research, 57,* 149-174.

Schunk, D. H. (1990). Goal-setting and self-efficacy during self-regulated learning. *Educational Psychologist, 25*, 71-86.

Schunk, D. H. (2012). Social cognitive theory. In D. Schunk (Ed.), *Learning theories: An educational perspective* (6th ed., pp. 117-162). Boston, MA: Pearson.

Searle, S. R. (1987). *Linear models for unbalanced data.* New York: John Wiley.

Semke, H. (1984). Effects of the red pen. *Foreign Language Annals, 17*, 195-202.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin Company.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Shook, S. E., Marrion, L. V., & Ollila, L. O. (1989). Primary children's concepts about writing. *The Journal of Educational Research, 82*(3), 133-139.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189. doi:10.3102/0034654307313795

Smith, S. (1997). The genre of the end comment: Conventions in teacher responses to student writing. *College Composition and Communication, 48*, 249-268.

Smits, M. (2009). *Schrijven en leren op de pabo.* Nijmegen: Proefschrift Radboud Universiteit.

Solomon, S. R. & Sawilowsky, S. S. (2009). Impact of rank-based normalizing trans formations on the accuracy of test scores. *Journal of Modern Applied Statistical Methods, 8*, 448-462.

Sommers, N. (1980). Revision strategies of student writers and experienced adult writers. *College Composition and Communication, 31*, 378-388.

Sommers, N. (1982). Responding to student writing. *College Composition and Communication, 33*, 148–156.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9.* Retrieved from http://PAREonline.net/getvn.asp?v=9&n=4

Stern, L. A. & Solomon, A. (2006). Effective faculty feedback: The road less traveled. *Assessing Writing, 11*, 22-41.

Stoeldraijer, J. (2012). *Kwaliteitskaart onderwijs in het schrijven van teksten.* Den Haag: School aan Zet.

Straub, R. (1996). The Concept of Control in Teacher Response: Defining the Varieties of "Directive" and "Facilitative" Commentary. *College Composition and Communication, 47*, 223–251.

Sun, M., Penuel, W. R., Frank, K. A., Gallagher, H. A., & Youngs, P. (2013). Shaping professional development to promote the diffusion of instructional expertise among teachers. *Educational Evaluation and Policy Analysis, 35*, 344-369. doi:10.3102/0162373713482763.

Suppes, P., & Zinnes, J.L. (1963). Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 39-74). New York: John Wiley & Sons.

Swales, J. (1990). *Genre analysis.* Cambridge: Cambridge University Press.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Tillema, M. (2012). *Writing in first and second language: Empirical studies on text quality and writing processes* (Unpublished doctoral dissertation). Utrecht: LOT

Dissertation Series.

Tillema, M., Van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2012). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing, 30*, 71-97. doi:10.1177/0265532212442647

Torrance, M., Fidalgo, R., & Robledo, P. (2015). Do sixth-grade writers need process strategies? *British Journal of Educational Psychology, 85*(1), 91-112. doi:10.1111/bjep.12065

Tracy, B., Reid, R., & Graham, S. (2009). Teaching young students strategies for planning and drafting stories: The impact of self-regulated strategy development. *The Journal of Educational Research, 102*(5), 323-332.

Trapman, M., Van Steensel, R., Van Schooten, E., Van Gelderen, A., & Hulstijn, J. (2012). Een longitudinale studie naar de rol van linguistische kennis, vloeiendheid en metacognitieve kennis in schrijfvaardigheid van leerlingen in het vmbo [A longitudinal study of the role of linguistic knowledge, fluency, and meta cognitive knowledge of students in junior intermediate vocational education]. In: N. van Jong, K. Juffermans, M. Keijzer, & L. Rasier (Eds), *Papers of the Anéla 2012: Applied Linguistics Conference* (pp 66-74). Delft: Eburon.

Troia, G., & Graham, S. (2002). The effectiveness of a highly explicit, teacher-directed strategy instruction routine: Changing the writing performance of students with learning disabilities. *Journal of Learning Disabilities, 35*, 290-305. doi:10.1177/00222194020350040101

Tschannen-Moran, M., Hoy, A. W., & Hoy, W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*, 202-248.

Underwood, J. S., & Tregidgo, A. P. (2010). Improving Student Writing Through Effective Feedback: Best Practices and Recommendations. *Journal of Teaching Writing, 22*, 73–98.

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

Van den Bergh, H., Maeyer, S. de, Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. In E. Van Steendam, M. Tillema, G. Rijlaarsdam, & H. Van den Bergh (Eds.), *Measuring writing: Recent insights into theory, methodology and practice* (Vol. 27, pp. 23–32). Leiden: Brill.

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

Van den Bergh, L., Ros, A., & Beijaard, D., (2013). Feedback during active learning: Elementary school teachers' beliefs and perceived problems. *Educational Studies, 39*, 418-430. doi:10.1080/03055698.2013.767188

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

Van der Leeuw, B. (2006). *Schrijftaken in de lerarenopleiding: een etnografie van onderwijsvernieuwing [Written assignments in teacher training: An ethnography of educational reform*] (Unpublished doctoral dissertation). Utrecht University.

Van der Leeuw, B., Pauw, I., Smits, M. & Van de Ven, P., (2010). Schrijven op de pabo; wat weten we uit onderzoek? [Writing in teaching training college; what does research tell us?]. *Levende Talen Tijdschrift, 11*(2), 30-39.

Van Steendam, E., Rijlaarsdam, G., Van den Bergh, H., & Sercu, L (2014). The mediating effect of instruction on pair composition in L2 revision and writing. *Instructional Science, 42*, 905-927. doi:10.1007/s11251-014-9318-5

Van Weijen, D. (2009). Writing processes, text quality, and task effects; empirical studies in first and second language writing (Unpublished doctoral dissertation). Utrecht: LOT Dissertation Series.

Veal, L. R., & Tillman, M. (1971). Mode of discourse variation in the evaluation of children's writing. *Research in the Teaching of English, 5*(1), 37–45.

Verheyden, L. (2011). *Achter de lijn. Vier empirische studies over ontluikende stelvaardigheid* (Unpublished doctoral dissertation). Leuven, Belgium: Katholieke Universiteit Leuven.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes.* Cambridge, MA: Harvard University Press.

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

Wesdorp, H. (1981). Evaluatietechnieken voor het moedertaalonderwijs [Evaluation techniques for the mother tongue education]. The Hague: SVO.

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation, 37*, 3–14. doi:10.1016/j.stueduc.2011.03.001

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing, 17,* 150–173. doi:10.1016/j.asw.2011.12.001

Wong, B. Y., Hoskyn, M., Jai, D., Ellis, P., & Watson, K. (2008). The comparative efficacy of two approaches to teaching sixth graders opinion essay writing. *Contemporary Educational Psychology, 33*, 757-784. doi:10.1016/j.cedpsych. 2007.12.004

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17,* 89-100.

Zamel, V. (1985). Responding to student writing. *TESOL Quarterly, 19,* 80–102.

Zellermayer, M. (1989). The study of teachers" written feedback to students" writing: changes in theoretical considerations and the expansion of research contexts. *Instructional Science, 18*, 145–165. doi:10.1007/BF00117715

Zhang, L., & Vukelich, C. (1998). *Prewriting Activities and Gender: Influences on the Writing Quality of Male and Female Students.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

Zimmerman, B. J., & Kitsantas, A. (2002). Acquiring writing revision and self-regulatory skill through observation and emulation. *Journal of Educational Psychology, 94*, 660-668. doi:10.1037/0022-0663.94.4.660

Zimmerman, B. J., & Risemberg, R. (1997). Becoming a self-regulated writer: A social cognitive perspective. *Contemporary Educational Psychology, 22*, 73-101. doi:10.1006/ceps.1997.0919

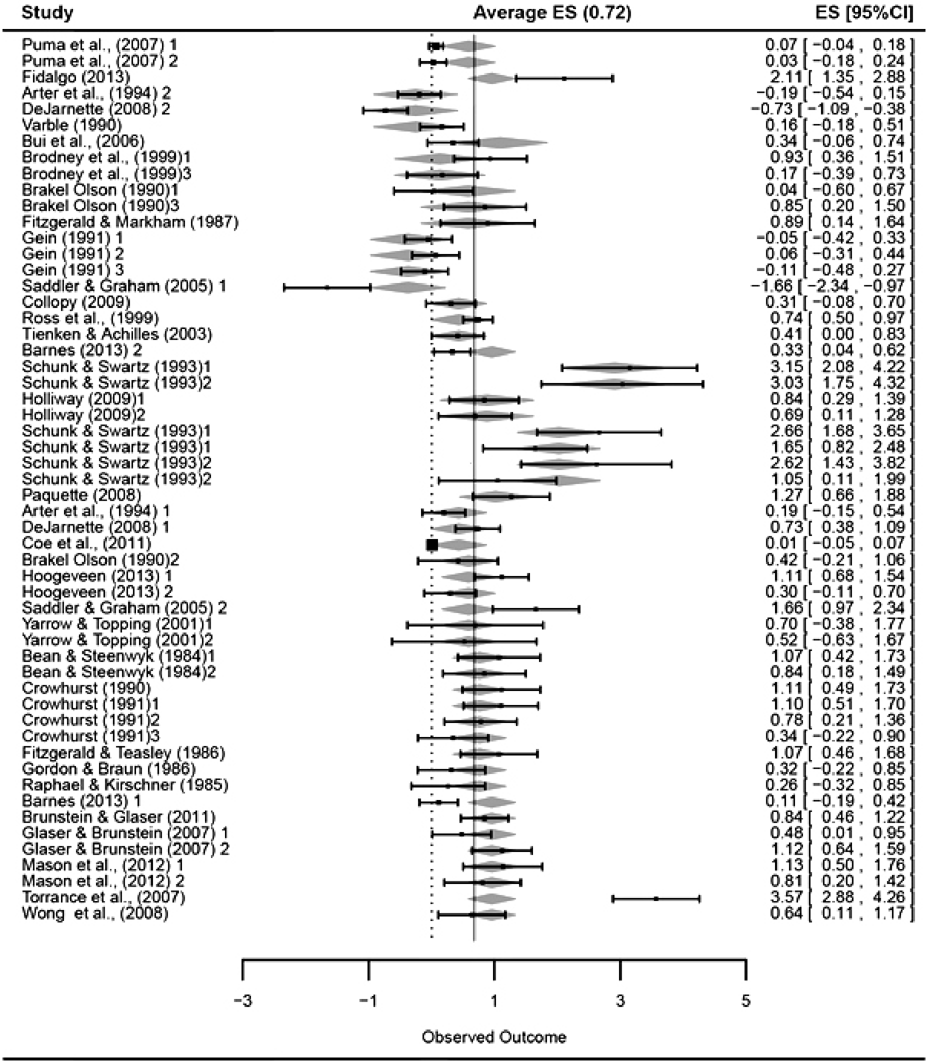
Ziv, N. D. (1980). *The effect of teacher comments on the writing of four college freshmen* (No. ED 203317). Retrieved from Eric Document Reproduction Service: http://eric.ed.gov

References

OVERVIEW OF APPENDICES

APPENDIX A

Forest plot of expected (grey) and observed effect sizes, including 95% confidence interval, per study in the meta-analysis (Chapter 2).



APPENDIX B

An overview of the argumentative and narrative writing tasks that are used in the generalizability study (Chapter 3). The 12 writing tasks are categorized according to purpose of writing and audience specification.

	Audience	
Purpose	**Specified reader**	**Unspecified reader**
Argumentative writing	Persuasive letters for a fictional company Topic 1: Collection of toys at a supermarket Topic 2: Collection of stamps at a petrol station for earning musical tickets Topic 3: Collection of points on wraps of chocolate bars to earn a music CD	Argumentative essays to prepare oneself for a class discussion Topic 1: Pros and cons of a candy prohibition for children Topic 2: Pros and cons of a smoking ban Topic 3: Pros and cons of telling tales about somebody
Narrative writing	Adventure stories for readers of a school newspaper Topic 1: Adventure on a sports field Topic 2: Adventure about a forest-fire Topic 3: Adventure about poison	Personal stories Topic 1: Personal experience about being frightened by something Topic 2: Personal experience about being caught for something Topic 3: Personal experience about being home alone

APPENDIX C

Task for writing a persuasive letter to a fictional company about saving Yummy points, used in Chapter 4.

Imagine … You are a real fan of the Yummy Yummy Candy Bars. One day you read the following advertisement:

> SAVE UP FOR A FREE MUSIC CD!
> How to get them:
> On the wrapper of each Yummy Yummy Candy Bar you will find 1 point. Save 10 points. Send the points in a sufficiently stamped envelope to:
>
> Yummy Yummy Promotion Campaign
> PO Box 3333
> 1273 AD Etten-Leur
>
> Include a stamp of 80 cents for the postage costs. Mention clearly your name, address, and the zip code of your residence. The free (FREE!) music CD will be sent as soon as possible to your address.
> This offer ends on July 15.

Now you have saved 8 points. Nearly all 10 points required! But you cannot find any more Yummy Yummy Bars with points on the wrapper, although it isn't July 15 yet. You tried different shops. So it seems you can't collect your 10 points. But you still want to get your free music CD. Therefore you decide to send your 8 points and 2 Yummy Yummy wrappers without points.

Task
Write a letter that you send with the 8 points and the 2 wrappers. Explain why you cannot send 10 points. Convince the Yummy Yummy Company that it isn't your fault that you didn't collect 10 points and that you still want to receive the music CD. Be sure they will send you the CD! Then write an envelope.

APPENDIX D

Task for writing a persuasive letter to a fictional company about the collection of Smurfs, used in Chapter 4.

Imagine … your dad does his weekly grocery shopping at the supermarket (Supercoop) around the corner. For every twenty-five euro he spends, he receives a mini-Smurf (packed in a bag). You have quite a collection already, but you're still missing two Smurfs: Brainy Smurf and Papa Smurf. Of course you're also hoping to get the Golden Smurf. If you get the Golden Smurf, you have the chance of winning a digital camera. Every Friday you look in the grocery bag with great anticipation to see which Smurfs you can add to your collection. One day you read the following advertisement in the folder of the supermarket:

> SAVE UP FOR SMURFS AND WIN A DIGITAL CAMERA!
> How to get them:
> For every 25 euro on groceries, you will receive a Smurf. This way you can surprise your kids with a wonderful collection of these popular little characters.
> Have you received a Golden Smurf, than you have a chance of winning a digital camera with 4.0 megapixels!
> Send your Golden Smurf in a sufficiently stamped envelope to:
> Smurf Promotion Campaign
> PO Box 3333
> 1273 AD Etten-Leur
> Include a stamp of 80 cents for the postage costs. Mention clearly your name, address, and the zip code of your residence.
> This offer ends on April 20. You will be informed about winning the camera before May 1.

On April 15 your dad tells you the bags of Smurfs ran out. He received a receipt with a stamp, though, so you can collect the Smurfs at a later time. He spent a total of 110,34 euro's on groceries. You decide to send the receipt, asking to receive the bags of Smurfs before April 20. Of course you're hoping to get the Golden Smurf.

Task
Write a letter that you send with the receipt of April 15. Explain that your local Supercoop ran out of Smurfs. Convince the Supercoop Company that you want to receive the Smurfs before the closing date of the campaign, so that you have a chance of winning the digital camera. Make sure they will send you the Smurfs! Then write an envelope.

APPENDIX E

Task for writing a letter of advice to Like, a fictional peer. This writing task is used in Chapter 4 and 6.

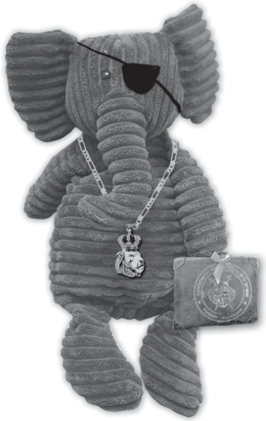
> Next week, a new student will arrive in your classroom: Like. Like was born in the Netherlands, but has lived in England for some time. The school system in England is different from the Dutch system. Like is not exactly sure how to write a good text in Dutch, because there are many aspects that you have to take into account. Write a letter to Like to explain how to write a good text in Dutch, and thus, to get good grades for writing. Give Like as many tips and advice you can think of.

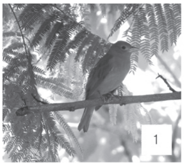
APPENDIX F

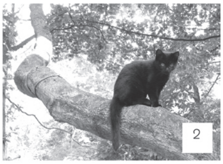
Examples of descriptive, narrative, and persuasive writing prompts, used in Chapter 5.

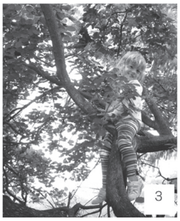
Descriptive writing prompt: Lost cuddly toy
Your little brother has lost his cuddly toy in the train. He is very sad and he desperately wants it back. That is why you want to put a message on the website of the 'Lost Property Department'.

Write a message in which you describe what the cuddly toy looked liked, and where and when your brother lost it. Remember to mention your name and address to make sure that the finder can contact you.









Narrative writing prompt: Cat in a tree
You see 3 pictures. They are the beginning of an exciting story about a cat in a tree. How will this story end? Make up the ending of the story.

Write down the story from beginning to end, and also think of a good title for your story.

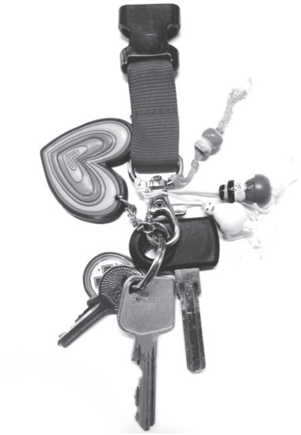
Persuasive writing prompt: Classroom pet
You and your classmates want a classroom pet. Your teacher does not find this a good idea. Still you want to do everything you can to try to get this pet.

Write a letter to your teacher in which you try to convince him or her with good arguments to get you a classroom pet. Clearly state in your letter what kind of pet you want.

APPENDIX G

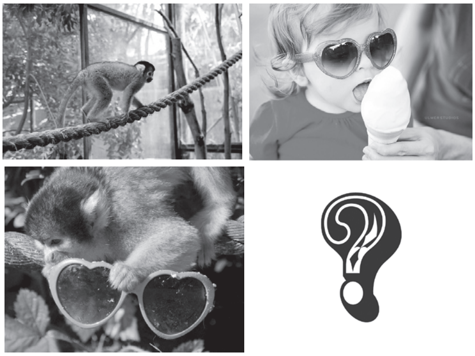
Examples of descriptive, narrative, and persuasive writing prompts, used in Chapter 7 and 9.

Descriptive writing prompt: Lost keyring
While she was shopping in the supermarket, your mother has lost her keyring. She is very sad and she desperately wants it back. You want to help her by putting a notice on the notice board in the supermarket.



Write a notice in which you ask whether anyone has found the keys. Describe what the keyring looks like, and where and when your mother lost it. Remember to mention your name and address to make sure that the finder can contact you.

Narrative writing prompt: Monkey
You see 3 pictures. They are the beginning of an exciting story about a monkey in the zoo. How will this story end? Make up the ending of the story.



Write down the story from beginning to end, and also think of a good title for your story.

Persuasive writing prompt: Amusement park
You and your classmates want a make a daytrip to an amusement park. Your teacher does not find this a good idea. Still, you want to do everything you can to try to make this happen.

Write a letter to your teacher in which you try to convince him or her with good arguments to go on a daytrip to an amusement park with the whole class. Clearly state in your letter to which amusement park you want to go.

APPENDIX H

Analytic scoring guide for the Yummy writing task, used in Chapter 4.

Dimension 1: Content

This concerns rating the inclusion of given and self-generated content elements.

CON1. Does the student mention that he has saved up eight points?

CON2. Does the student mention that he has enclosed two wrappers without points?

CON3. Does the student mention that the campaign has not ended yet?

CON4. Does the student argue why he has a right to receive the cd?

CON5. Does the student add extra elements to support his argument?

CON6. Does the letter include enough information about address or other information to enable deliverance of the cd?

Dimension 2: Organization and structure

This concerns the organization of the content and the structure of the text.

STR1. Does the letter contain a formal salutation?

STR2. Does the student start his letter with words that are intended as an introduction?

STR3. Does the student present information in a logical order?

STR4. Does the letter contain words that are intended as a closing?

STR5. Does the letter end with a more or less conventional valediction or closing?

STR6. Is the letter signed?

Dimension 3: Communication

This concerns focus on goal and audience.

COM1. Is the argumentation sufficiently convincing?

COM2. Does the student clarify what he expects from the reader?

COM3. Does the student address the reader?

APPENDIX I

Benchmark rating scale for persuasive letters, used in Chapter 4. Benchmarks on the scale are from the Yummy writing task, see Appendix C.

I have a question for yummy.
Now I have 8 points but in all shops I can't get any
more and
it is not July 15 yet.
And I want a cd.
I want too ask you if I will still get a cd will you please let me know then I wil get it or
not.

Greetings.

70

This text is even worse because:

- The problem isn't clarified enough. It isn't clear that it's about a promotion campaign and that 2 wrappers without points are enclosed.
- No address has been added, so the cd can't be sent.
- The structure is poor, information isn't presented in a logical order.
- It doesn't follow letter conventions, there is no salutation or signature.
- The letter's grammar and punctuation aren't correct (spelling is ok).

Dear Yummy Yummy Company
I want the CD veryr badly. I have saved for it as well but
sadly I
have only 8th coins that's not
enough.
but I'm sending wrappers with it.
Will you please think about it.
Thanks in advance

<name + adress>

85

This text is worse because:

- The problem isn't clarified enough. It's not clear that the campaign hasn't ended and that the wrappers ran out, causing the writer to have too few points.
- The question hasn't been asked very clearly. The reader isn't being convinced to actually send the cd.
- The letter contains errors in grammar (e.g. eight*h* points) and isn't flawless regarding spelling and punctuation either.

Dear Sir/Madam ov the yummy campaign we have saved 8 points but now we can't find
any more points
can but it isn't July 15 couldn you
make an exception for that. Since
it's only two points would you send the
cd anyway because we looked on every
shelf if they were attached but no
found nothing. the address is <student's
address>

With regards from <name>

100

This is an average text because:

- The problem is more or less described: the writer hasn't been able to collect enough Yummy points, while the campaign hasn't ended yet.
- A clear question is asked.
- The sender's address is mentioned.
- The letter has a conventional salutation and closing.

Weaknesses:
- The enclosure of two separate wrappers isn't mentioned.
- The student uses long sentences that are difficult to follow because of their weak (grammatical) structure.
- Grammar, spelling, punctuation and layout are weak.

Dear Sir,

I took part in the promotion campaign and have collected 8 points, it's not July 15 yet and there are no points on the wrappers anymore, while there are still enough left.
I bought bars and am sending 2 without points. Could I please also have the cd?
I really would like one and there were none left on them. So I am could not reach 10 points. I would be very happy if you could send it.
Because I can keep looking for these bars
I would find many: but without points.

With kind regards

< name + adress>

115

This text is better because:

- The problem is described clearly: writer hasn't collected enough points and there are no more points available in the store, while the campaign is still going; it's clear that that's why two wrappers have been sent.
- A clear question is asked.
- The formal organization of the letter is correct: contains good salutation + closing, address is mentioned.
- Spelling and punctuation are fine.

Some weaknesses:
- The structure isn't that good. Halfway the writer repeats himself and the message becomes unclear, partly because of a grammatical mistake.
- The tone of the request could have been more appropriate

Breda June 09

Dear Yummy Yummy Company,

When I read about the campaign I started collecting right away. I had 8 points already and the campaign hasn't ended yet.
I reeeeeeeally want the cd!
But the points have run out.
I have looked on every wrapper in as many stores as possible, even on the other side of the country! but sadly they really ran out.
I have added two normals wrappers to prove that I should have ten. (I added 80 cents as well)
I hope to receive one anyway.

With kind regards,

<address>

130

This text is even better because:

- The letter is very convincing, the problem is described adequately and the question is asked clearly.
- The structure is good, making the content comprehensible.
- The composition of the letter is good: formal salutation + closing, address is mentioned.
- The text is grammatically correct, spelling and punctuation are satisfactory.

APPENDIX J

Benchmark rating scale for letters of advice, used in Chapter 4 and 6. Benchmarks on the scale are from the Like writing task, see Appendix E.

Dear Like,
do you have friends do you also have a hoby ~~what is~~ what kin of thins do you do do you also play sport. do you have a pet. do you walk outside often. ~~what do you eat often~~ what do you like to eat. what is your favorite drink.

70

70 points
Strengths:
- good salutation

Weaknesses:
- letter does not meet the assignment: no tips are given on how to write a good text
- no concluding sentence, letter is not signed
- punctuation: capitals and question marks are missing
- spelling mistakes

Dear Like,
Start a sentence with a capital and a full stop at the end. Write ~~neetly~~ neatly between the lines. Write a capital at the beginning of a name.
end

85

85 points

Strengths:
- good salutation
- writer gives three tips for writing a good text
- correct punctuation

Weaknesses:
- tips are superficial
- letter is merely enumeration of tips, lacking an introduction or motivation: why this letter?
- no concluding sentence, letter is not signed

Dear Like,
You need a well subject and you need to write neatly and think about the capitals. That way you can get good grades.
You have to pay atention to many things. The ~~the~~ most important is that you have an good subject. And think of the spelling rules.
Good luck, Manon

100

100 points
Strengths:
- good salutation
- writer gives four tips
- writer provides extra information: you have to pay attention to many things, the most important thing is..., etc.
- proper concluding sentence and signature
- correct punctuation

Weaknesses:
- tips are superficial
- letter contains repetition (2x good subject)
- spelling mistakes

Dear Like,
I am Abigail. I can help you with your text. I will some tips and for example:
always start with a capital when you begin a sentence and always end with a full stop. You should not make very long sentences but nice shorter sentences. But if you really have to make one long thehave to put a comma in the sentence. If you don't know how to write something then write down how you thik think it should be written.
I wish you lots of luck with your tekst.
Lots of love from Abigail

115

115 points

Strengths:
- good salutation
- writer gives many tips and provides explanation
- good introduction: writers introduces herself and starts with the goal of the letter
- proper concluding sentence and signature
- letter has a friendly tone
- writer uses blank lines
- no spelling mistakes

Weaknesses:
- writer uses long

Dear Like,
How nice that you come to the Netherlands.
Writing in Dutch can be difficult.
It is different from England.
Dutch is a hard language (you probably know already).
So I will give you some tips.
Tip 1: Think well what you want to write do not write down just anything
Tip2: Spelling mistakes can happen do no get nervous it happens often
Tip3: If you struggle ask the teacher or grab a piece of paper and write down the words you are unsure of, handy isn't it?
Off the record: do not cheat, that can end like this:
So, you are cheating. Turn it in and the teacher marks it. And sees at once that you have exactly the same answers. She want to have a serious talk, she phones your parents to tell them, and guess what. You are banned from the computer for a week.
Ouch. Not funny that is what happened to me too.
What I want to say to you is that you must never cheat or throw pellets.
So in this case cheating 'Watch out'.
Tip 4: You can always look in the dictionary during language or spelling.
Tip 5: You can also clap the words like: paper – pa- per.
Tip 6: Use exclamation marks (!) or question marks (?) often. Very tampting for teachers.
And don't think that you cannot do it it will be all right, you just have to be patient. These were my tips. More next time ☺!
Lots of luck xx
Brownie

130

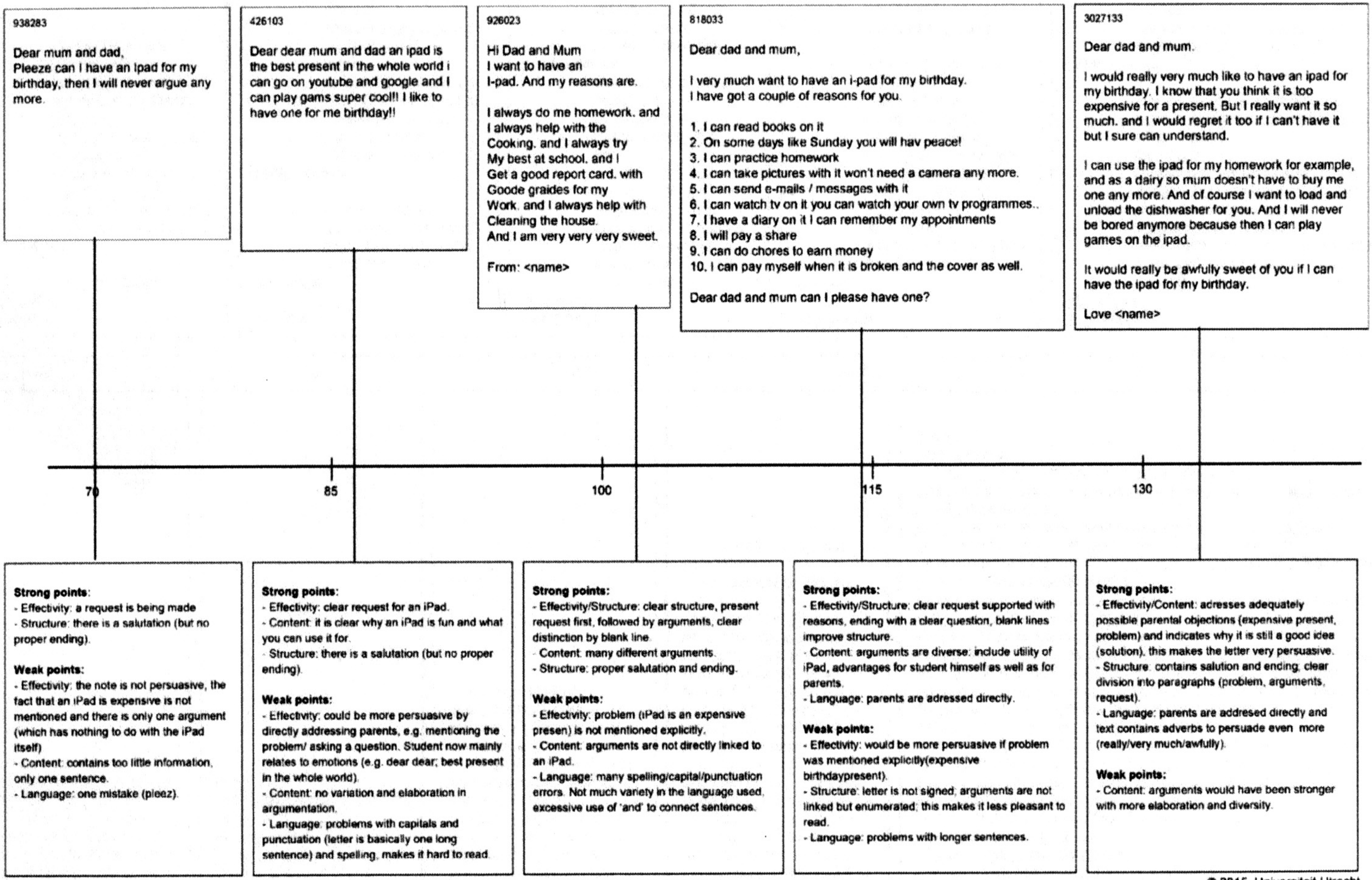
130 points
Strengths:
- good salutation
- good introduction: writer gives motivation for the letter and directly addresses the reader
- writer gives many clear
- the tips are clearly indicated (Tip1:)
- proper concluding sentence and signature
- style is lively
- letter contains humor (ouch, throw pellets, tempting for teachers)

Weaknesses:
- writer uses long sentences
- punctuation errors: lacking commas and full stops
- spelling mistake

APPENDIX K

A benchmark rating scale for persuasive letters, used in Chapter 5, 7, and 9. Benchmarks on the scale are from the iPad writing task, in which students are prompted to write a letter to their parents in order to convince them to give an iPad for their birthday.

Rating scale for persuasive essays



938283

Dear mum and dad,
Pleeze can I have an ipad for my birthday, then I will never argue any more.

Strong points:
- Effectivity: a request is being made
- Structure: there is a salutation (but no proper ending).

Weak points:
- Effectivity: the note is not persuasive, the fact that an iPad is expensive is not mentioned and there is only one argument (which has nothing to do with the iPad itself)
- Content: contains too little information, only one sentence.
- Language: one mistake (pleez).

426103

Dear dear mum and dad an ipad is the best present in the whole world i can go on youtube and google and I can play gams super cool!! I like to have one for me birthday!!

Strong points:
- Effectivity: clear request for an iPad.
- Content: it is clear why an iPad is fun and what you can use it for.
- Structure: there is a salutation (but no proper ending).

Weak points:
- Effectivity: could be more persuasive by directly addressing parents, e.g. mentioning the problem/ asking a question. Student now mainly relates to emotions (e.g. dear dear; best present in the whole world).
- Content: no variation and elaboration in argumentation.
- Language: problems with capitals and punctuation (letter is basically one long sentence) and spelling, makes it hard to read.

926023

Hi Dad and Mum
I want to have an
I-pad. And my reasons are.

I always do me homework. and
I always help with the
Cooking. and I always try
My best at school. and I
Get a good report card. with
Goode graides for my
Work. and I always help with
Cleaning the house.
And I am very very very sweet.

From: <name>

Strong points:
- Effectivity/Structure: clear structure, present request first, followed by arguments, clear distinction by blank line.
- Content: many different arguments.
- Structure: proper salutation and ending.

Weak points:
- Effectivity: problem (iPad is an expensive presen) is not mentioned explicitly.
- Content: arguments are not directly linked to an iPad.
- Language: many spelling/capital/punctuation errors. Not much variety in the language used, excessive use of 'and' to connect sentences.

818033

Dear dad and mum,

I very much want to have an i-pad for my birthday.
I have got a couple of reasons for you.

1. I can read books on it
2. On some days like Sunday you will hav peace!
3. I can practice homework
4. I can take pictures with it won't need a camera any more.
5. I can send e-mails / messages with it
6. I can watch tv on it you can watch your own tv programmes..
7. I have a diary on it I can remember my appointments
8. I will pay a share
9. I can do chores to earn money
10. I can pay myself when it is broken and the cover as well.

Dear dad and mum can I please have one?

Strong points:
- Effectivity/Structure: clear request supported with reasons, ending with a clear question, blank lines improve structure.
- Content: arguments are diverse: include utility of iPad, advantages for student himself as well as for parents.
- Language: parents are adressed directly.

Weak points:
- Effectivity: would be more persuasive if problem was mentioned explicitly(expensive birthdaypresent).
- Structure: letter is not signed; arguments are not linked but enumerated; this makes it less pleasant to read.
- Language: problems with longer sentences.

3027133

Dear dad and mum.

I would really very much like to have an ipad for my birthday. I know that you think it is too expensive for a present. But I really want it so much. and I would regret it too if I can't have it but I sure can understand.

I can use the ipad for my homework for example, and as a dairy so mum doesn't have to buy me one any more. And of course I want to load and unload the dishwasher for you. And I will never be bored anymore because then I can play games on the ipad.

It would really be awfully sweet of you if I can have the ipad for my birthday.

Love <name>

Strong points:
- Effectivity/Content: adresses adequately possible parental objections (expensive present, problem) and indicates why it is still a good idea (solution). this makes the letter very persuasive.
- Structure: contains salution and ending, clear division into paragraphs (problem, arguments, request).
- Language: parents are addresed directly and text contains adverbs to persuade even more (really/very much/awfully).

Weak points:
- Content: arguments would have been stronger with more elaboration and diversity.

APPENDIX L

Overview of rater reliability for the two intervention studies described in Chapter 5 and 9.

Table L1 Reliabilities of individual raters (*N* = 47) and juries (3 raters per jury, 47 juries) for intervention study 1

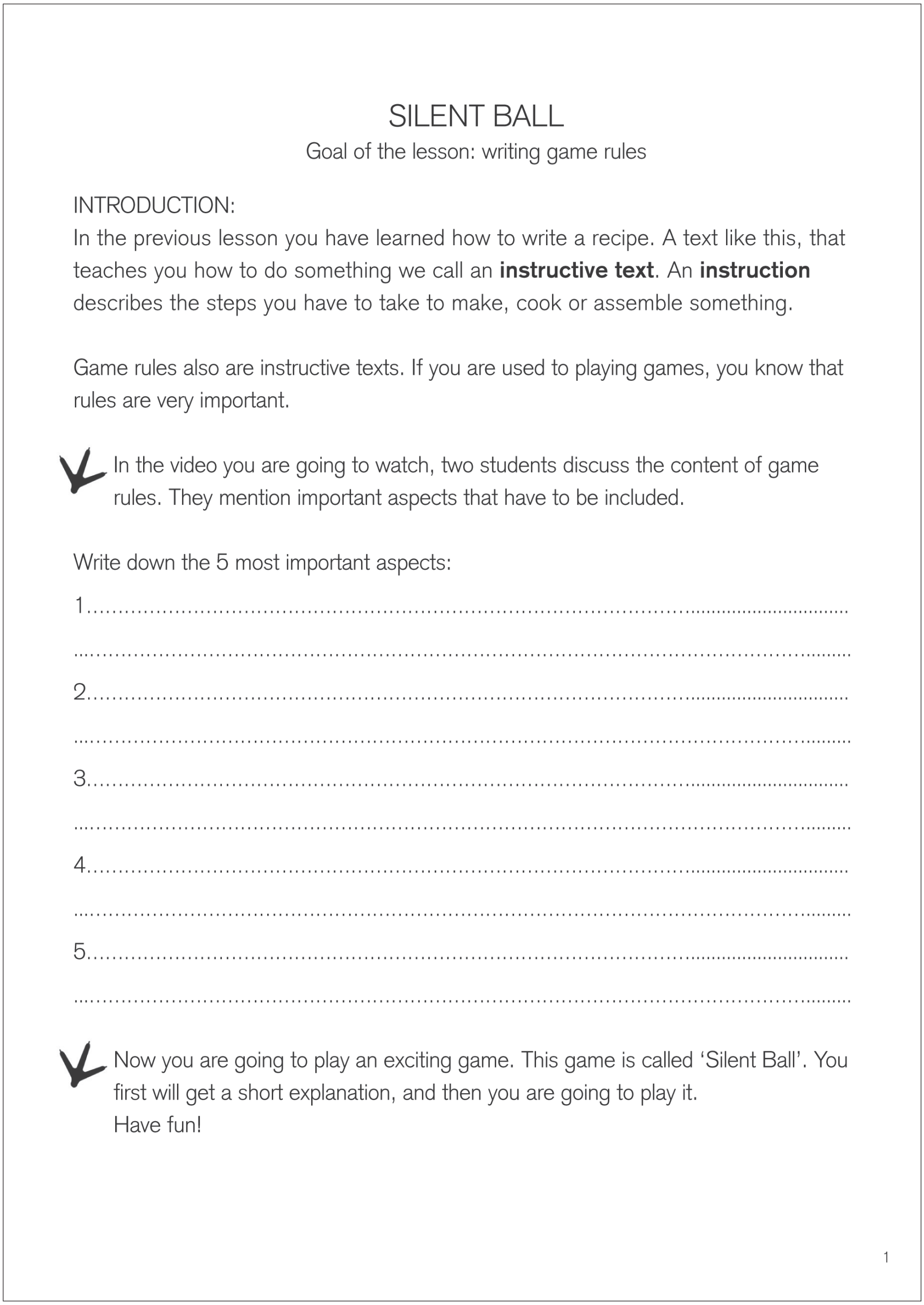
Genre	Measurement occasion	Task	Reliabilities of individual raters (mean *ρ*, *SD*)	Reliability of jury's of three raters (mean *ρ*, *SD*)
Descriptions	1	Child's bicycle	.69 (.14)	.87 (.04)
	2	Teddy bear	.68 (.15)	.86 (.04)
	3	Pickpocket	.72 (.14)	.88 (.04)
Narratives	1	Talent show	.77 (.11)	.91 (.03)
	2	Cat in a tree	.76 (.12)	.91 (.03)
	3	Bicycle competition	.75 (.14)	.89 (.04)
Persuasive letters	1	iPad	.78 (.11)	.91 (.02)
	2	Classroom pet	.74 (.13)	.89 (.03)
	3	Summer vacation	.70 (.13)	.87 (.03)
Average			.73 (.13)	.89 (.03)

Table L2 Reliabilities of individual raters (*N* = 18) and juries (3 raters per jury, 18 juries) for intervention study 2

Genre	Measurement occasion	Task	Reliabilities of individual raters (mean *ρ*, *SD*)	Reliability of jury's of three raters (mean *ρ*, *SD*)
Descriptions	1	Lost keyring	.69 (.12)	.87 (.03)
	2	Child's bicycle	.69 (.07)	.87 (.02)
	3	Inflatable crocodile	.62 (.10)	.83 (.03)
Narratives	1	Talent show	.75 (.07)	.90 (.02)
	2	Monkey	.71 (.10)	.88 (.02)
	3	Camping	.76 (.10)	.90 (.02)
Persuasive letters	1	Amusement park	.73 (.06)	.89 (.01)
	2	Play equipment	.72 (.08)	.88 (.02)
	3	iPad	.71 (.11)	.88 (.03)
Average			.71 (.09)	.88 (.02)

APPENDIX M

An example lesson of the teaching program Tekster, translated from Dutch.



SILENT BALL

Goal of the lesson: writing game rules

INTRODUCTION:
In the previous lesson you have learned how to write a recipe. A text like this, that teaches you how to do something we call an **instructive text**. An **instruction** describes the steps you have to take to make, cook or assemble something.

Game rules also are instructive texts. If you are used to playing games, you know that rules are very important.

In the video you are going to watch, two students discuss the content of game rules. They mention important aspects that have to be included.

Write down the 5 most important aspects:

1..

....

2..

....

3..

....

4..

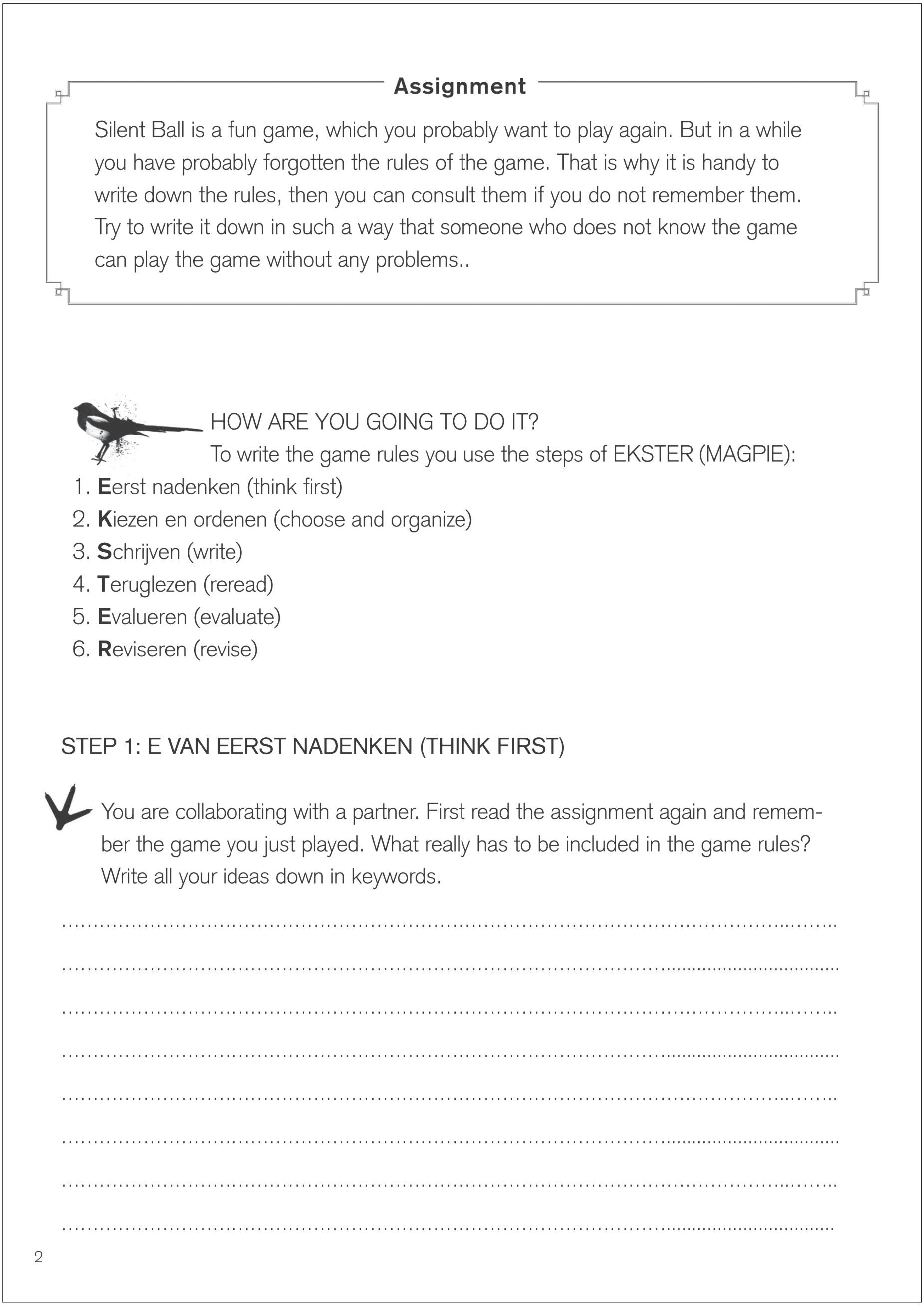
....

5..

....

Now you are going to play an exciting game. This game is called ‘Silent Ball’. You first will get a short explanation, and then you are going to play it.
Have fun!

1



Assignment

Silent Ball is a fun game, which you probably want to play again. But in a while you have probably forgotten the rules of the game. That is why it is handy to write down the rules, then you can consult them if you do not remember them. Try to write it down in such a way that someone who does not know the game can play the game without any problems..

HOW ARE YOU GOING TO DO IT?
To write the game rules you use the steps of EKSTER (MAGPIE):

1. **E**erst nadenken (think first)
2. **K**iezen en ordenen (choose and organize)
3. **S**chrijven (write)
4. **T**eruglezen (reread)
5. **E**valueren (evaluate)
6. **R**eviseren (revise)

STEP 1: E VAN EERST NADENKEN (THINK FIRST)

You are collaborating with a partner. First read the assignment again and remember the game you just played. What really has to be included in the game rules? Write all your ideas down in keywords.

……………………………………………………………………………………………………..……..

………………………………………………………………………………………………..................................

……………………………………………………………………………………………………..……..

………………………………………………………………………………………………..................................

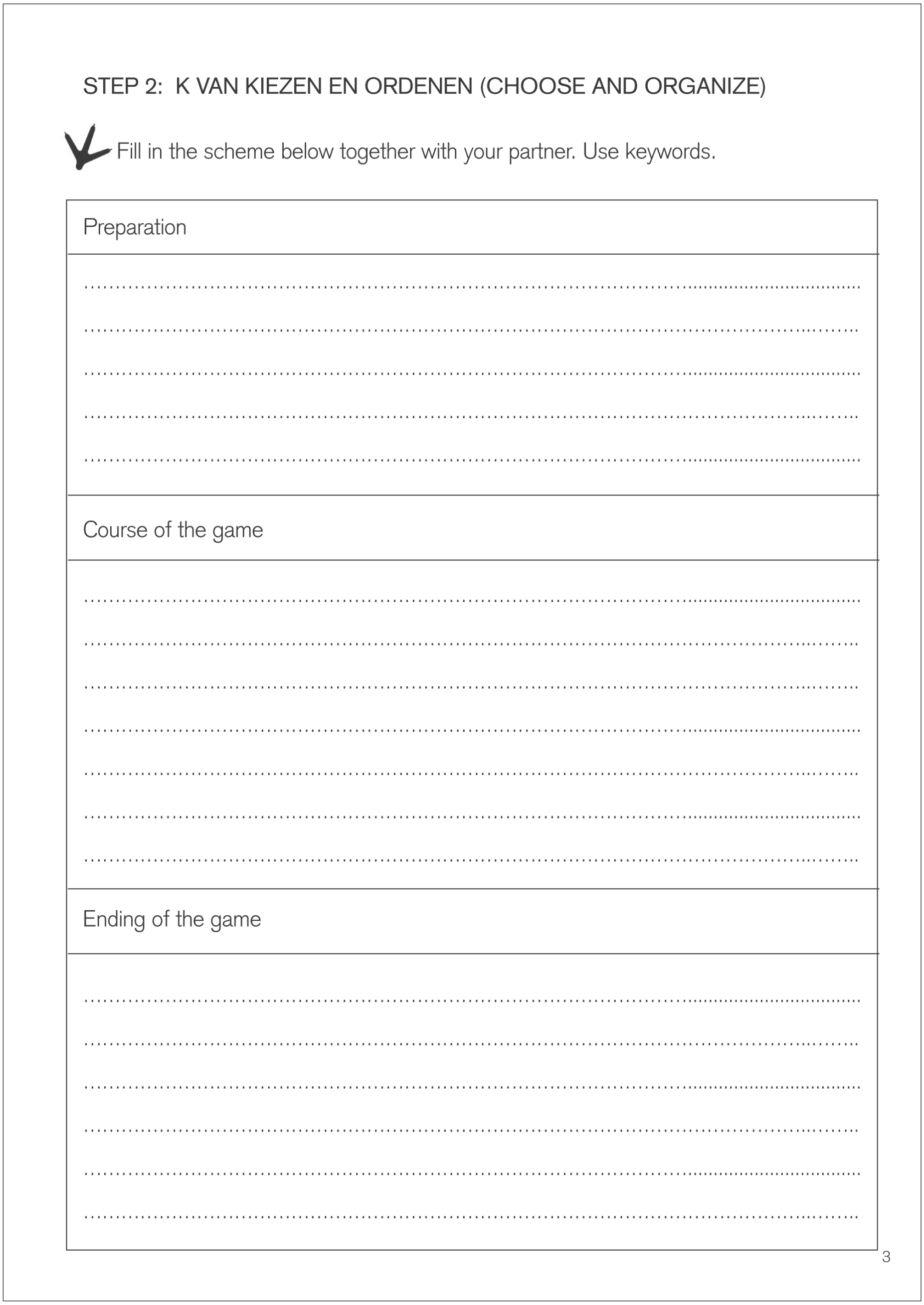
……………………………………………………………………………………………………..……..

………………………………………………………………………………………………..................................

……………………………………………………………………………………………………..……..

………………………………………………………………………………………………................................

2



STEP 2: K VAN KIEZEN EN ORDENEN (CHOOSE AND ORGANIZE)

Fill in the scheme below together with your partner. Use keywords.

Preparation

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

Course of the game

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

Ending of the game

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

3

STEP 3: S VAN SCHRIJVEN (WRITE)

You have thought about the rules of the game and the order of the rules. Now, write your rules down. Note that they must be clear for someone who is going to read them.

………

………

………

………

………

………

………

………

………

………

………

………

………

………

………

………

………

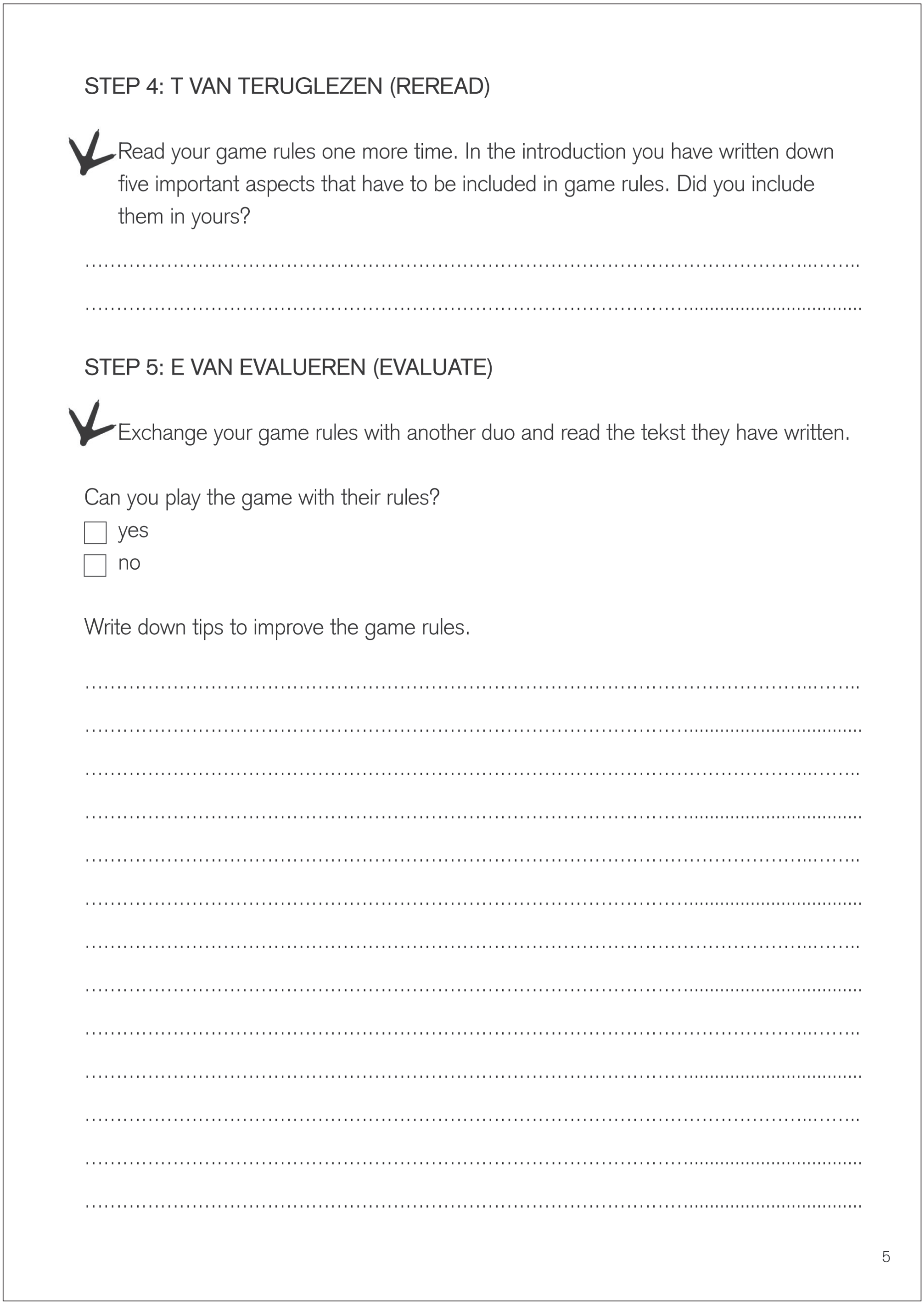
………

………

………

………

………



STEP 4: T VAN TERUGLEZEN (REREAD)

Read your game rules one more time. In the introduction you have written down five important aspects that have to be included in game rules. Did you include them in yours?

……………………………………………………………………………………………………

……………………………………………………………………………………………………

STEP 5: E VAN EVALUEREN (EVALUATE)

Exchange your game rules with another duo and read the tekst they have written.

Can you play the game with their rules?

□ yes
□ no

Write down tips to improve the game rules.

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

STEP 6: R VAN REVISEREN (REVISE)

Read the tips you have been given and revise your game rules. Write your revised version below.

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

Appendices

SCHRIJFONDERZOEK NAAR HET KLASLOKAAL

De effecten van Tekster, een nieuw ontwikkeld lesprogramma voor leerlingen in het basisonderwijs

Tijdens de vorige eeuw is onze maatschappij aanzienlijk veranderd. Waar voorheen de economie voornamelijk op landbouw en industrie was gericht, verdienen we in de huidige kenniseconomie ons geld voornamelijk door het verwerken en uitwisselen van informatie (Brandt, 1995). Hierdoor is de hoeveelheid schriftelijke informatie die we moeten produceren en verwerken enorm toegenomen. Een goede taalvaardigheid is van essentieel belang om volwaardig te kunnen deelnemen aan de huidige maatschappij. Verder heeft de opkomst van het internet en de snelle ontwikkelingen in communicatietechnologie de manier waarop we met elkaar communiceren grondig veranderd. We schrijven meer dan ooit tevoren, voor werk, maar ook privé, zoals bijvoorbeeld e-mail, berichten via sociale media, blogs, websites, en formulieren.

Een goede schrijfvaardigheid is dus essentieel voor beroepscarrière en deelname aan de maatschappij, maar het is tevens een van belangrijkste voorspellers voor schoolsucces (Graham & Perin, 2007). Schrijven is een basisvaardigheid, die helpt om gedachten te ordenen en het kritisch denken en logisch redeneren te stimuleren (Graham, 2013). Daarnaast wordt gedurende de schoolcarrière de meeste leerinhoud schriftelijk getoetst, waardoor betere schrijvers een grotere kans hebben om beter te presteren en hogerop te komen in de maatschappij. Het is daarom van essentieel belang dat leerlingen al vanaf jonge leeftijd goed leren schrijven. Het is een misvatting dat schrijven een aangeboren talent is waar je goed in bent of niet. Een andere populaire misvatting over schrijven is dat oefening kunst baart. Natuurlijk heb je oefening nodig om vooruitgang te boeken, maar je heb ook instructie en feedback nodig om te weten hoe je jezelf kunt verbeteren. In tegenstelling tot mondelinge taalvaardigheid, die je op een natuurlijke manier leert door blootstelling en taal en oefening vanaf de vroege levensjaren, is schrijven een vaardigheid die op school wordt aangeleerd, met behulp van instructie en begeleid oefenen.

Leren schrijven

Leren schrijven is niet makkelijk, zeker niet voor beginnende schrijvers. Schrijven is een complex cognitief proces waarbij een schrijver een flink aantal cognitieve activiteiten tegelijk moet uitvoeren, zoals bijvoorbeeld: wat wil ik zeggen, voor wie is de tekst bedoeld, hoe formuleer ik dit op een goede manier? Voor beginnende schrijvers is een bijkomend probleem dat handschrift, spelling en grammatica nog niet volledig geautomatiseerd zijn (McCutchen, 2011), waardoor deze aspecten nog een extra beroep doen op de cognitieve capaciteit van de schrijver. Bij beginnende schrijvers kan hierdoor cognitieve overbelasting ontstaan, waardoor ze terugvallen op een zogenoemde 'knowledge-telling' strategie:

ze schrijven op wat ze ter plekke bedenken, wat resulteert in 'en toen, en toen'-achtige verhalen (Bereiter & Scardamalia, 1987). Door cognitieve overbelasting hebben schrijvers vaak hun volledige aandacht nodig voor het produceren van een tekst en blijken zij amper toe te komen aan iets leren van hun taakuitvoering. Hiermee staat de schrijftaak los van het leerproces (Rijlaarsdam & Couzijn, 2000).

Huidige stand van zaken in het schrijfonderwijs

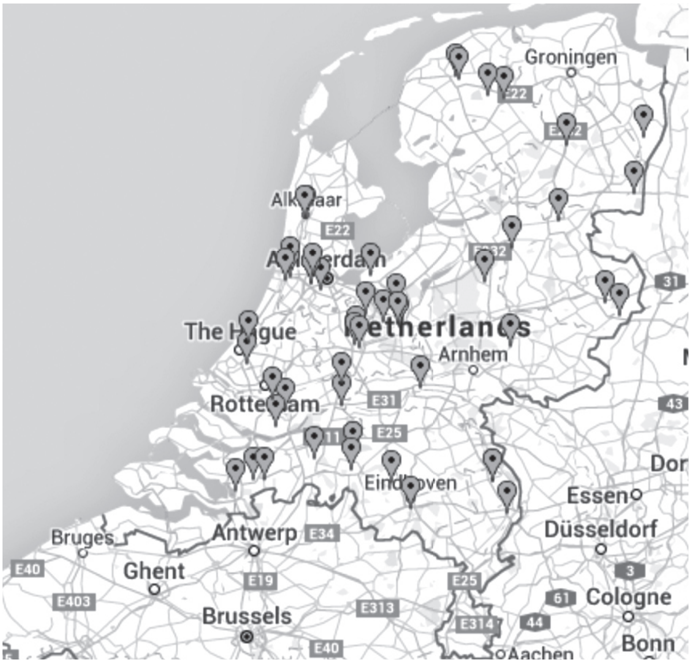
Het huidige schrijfonderwijs op de basisschool blijkt leerlingen onvoldoende te ondersteunen bij het leren schrijven. Uit de twee laatste grootschalige onderwijspeilingen in Nederland blijkt dat de schrijfprestaties van leerlingen ver onder het gewenste niveau zijn (Krom, Van de Gein, Van der Hoeven, Van der Schoot, Verhelst, Veldhuijzen & Hemker, 2004; Kuhlemeier, Van Til, Hemker, De Klijn, & Feenstra, 2013). Twee derde van de leerlingen is niet in staat om een simpele boodschap schriftelijk over te brengen aan een lezer. Verder blijken de schrijfprestaties van leerlingen nauwelijks vooruit te gaan tussen groep 6 en 8. Onderzoek van de Onderwijsinspectie (2010) heeft uitgewezen dat de gemiddelde kwaliteit van de schrijflessen in het basisonderwijs onvoldoende is. Er wordt weinig tijd aan het schrijven van teksten (stellen) besteed: van de 8 uur die gemiddeld per week voor taal op het lesrooster staan, wordt maar maximaal drie kwartier aan schrijfonderwijs besteed, vaak zonder enige vorm van instructie. De taalmethodes bieden docenten weinig houvast bij hun schrijfonderwijs: het aanbod voor schrijven is vaak onvoldoende en de docentenhandleidingen geven te weinig aanwijzingen voor het geven van schrijfinstructie (Henkens, 2010). Het huidige schrijfonderwijs is vooral productgeoriënteerd, met weinig aandacht voor het schrijfproces. Over het algemeen besteden docenten in hun schrijfinstructie weinig aandacht aan de aanpak van een schrijftaak, het bespreken van teksten, het geven van feedback en het herlezen en reviseren van teksten (Henkens, 2010). Ook op de lerarenopleidingen is de aandacht voor het schrijven beperkt (Van der Leeuw, 2007). Ondanks dat aankomende leraren zelf wel veel moeten schrijven voor hun studie, zoals portfolio's en reflectieverslagen, krijgen zij hier nauwelijks instructie in of feedback op en wordt er maar minimaal aandacht besteed aan schrijfdidactiek (Van der Leeuw, 2007). Een bijkomend probleem is dat er geen goede instrumenten beschikbaar zijn voor docenten om de schrijfprestaties van leerlingen in kaart te brengen. Het beoordelen van schrijfproducten is tijdrovend en vaak ook uiterst subjectief. Hierdoor is het voor docenten lastig is om goed zicht te krijgen op het schrijfvaardigheidsniveau van hun leerlingen, waardoor het vrijwel onmogelijk is om de schrijfontwikkeling te monitoren of bij te sturen over tijd.

Verbeteren van het schrijfonderwijs met Tekster

De voorgaande analyse laat zien dat het schrijfonderwijs op meerdere punten verbeterd kan worden: er is zowel behoefte aan beter lesmateriaal als aan instrumenten voor de toetsing en beoordeling van schrijven, als aan betere ondersteuning voor docenten om schrijflessen te kunnen geven. Om een impuls aan het schrijfonderwijs te geven hebben wij, in nauwe samenwerking met basisschooldocenten, Tekster ontwikkeld. Tekster is een strategie-gericht lesprogramma voor schrijfvaardigheid voor groep 6 t/m 8, gebaseerd op de bevindingen uit wetenschappelijk onderzoek naar effectieve schrijfdidactieken. Tekster bestaat uit drie lessenseries van 16 lessen, waarin leerlingen een aanpak leren voor het schrijven van teksten in verschillende genres. Verder bestaat de methode uit instrumenten voor toetsing en beoordeling van schrijfvaardigheid en een docentenhandleiding en training om de deskundigheid van docenten op het gebied van schrijfonderwijs te bevorderen. De effectiviteit van Tekster is beproefd in twee grootschalige interventiestudies waaraan in totaal 2766 leerlingen en 144 docenten deelnamen. Zij waren afkomstig van 52 scholen door heel Nederland, zie Figuur 1.

In deze samenvatting geven wij een overzicht van de resultaten van de studies die met Tekster zijn gedaan. Allereerst besteden we aandacht aan de wetenschappelijke onderbouwing van de keuzes die we hebben gemaakt bij het ontwikkelen van de lesmethode Tekster. Daarnaast zullen we de uitkomsten van de interventiestudies bespreken waarbij we zowel ingaan op de effectiviteit van het gehele programma als op de effectiviteit van specifieke componenten van Tekster. Ook zullen we bespreken wat de implicaties zijn van de onderzoeksresultaten voor het onderzoek naar schrijfvaardigheid en de onderwijspraktijk. Hierbij zullen we speciale aandacht besteden aan de implementatie van Tekster in het klaslokaal.

Figuur 1 Overzicht van de locatie van deelnemende scholen



DE BASISINGREDIËNTEN VAN HET LESPROGRAMMA TEKSTER

Effectieve didactische componenten voor leren schrijven

Om te achterhalen wat de meest effectieve manieren zijn voor het verbeteren van schrijfvaardigheid hebben we een meta-analyse uitgevoerd (zie hoofdstuk 2). In deze meta-analyse zijn de effecten vergeleken van interventiestudies die zich specifiek richten op de verbetering van schrijfvaardigheid van leerlingen van groep 6 tot en met 8 in het reguliere basisonderwijs. Onderzoek gericht op andere leeftijdsgroepen of op specifieke groepen leerlingen, zoals bijvoorbeeld leerlingen met leerproblemen of hoogbegaafdheid, is buiten beschouwing gelaten aangezien deze leerlingen mogelijk een andere behoefte aan ondersteuning nodig hebben dan een leerling in het regulier onderwijs. De resultaten van de meta-analyse laten zien dat vijf didactieken de schrijfprestaties significant verbeteren: doelen stellen (ES = 2.03), strategie-instructie (ES = 0.96), feedback (ES = 0.88), tekststructuurinstructie (ES = 0.76), en peer-interactie (ES = 0.59). Samen vormen deze vijf didactieken de basisingrediënten voor de lessen van Tekster.

Wat maakt deze didactieken zo effectief voor het leren schrijven? Het stellen van doelen is een belangrijk aspect van het schrijfproces. Je schrijft omdat je een boodschap wil overbrengen aan een lezer. Wat je wilt vertellen en aan wie bepaalt al voor een groot gedeelte hoe je tekst eruit zal gaan zien en welke taal je gebruikt. Een routebeschrijving ziet er anders uit dan een verhaal, en in een brief aan de burgemeester gebruik je andere taal dan in een kaartje aan je oma. Door het stellen van doelen leren leerlingen zelf de kwaliteit van hun tekst monitoren, waarbij de vraag centraal staat: behaal ik met deze tekst het gewenste doel bij de lezer? Door het aanleren van deze belangrijke zelfregulerende vaardigheden leert een leerling hoe hij zijn schrijfproces kan optimaliseren, wat leidt tot betere teksten (Schunk & Swartz, 1993). Een procesgerichte schrijfstrategie kan leerlingen nog verder ondersteunen tijdens het schrijfproces (o.a. Brunstein & Glaser, 2011; Torrance, Fidalgo, & Garcia, 2007). Door het opdelen van het schrijfproces in kleinere stappen vermindert de cognitieve overbelasting tijdens het schrijven. Als leerlingen voor het schrijven eerst ideeën genereren en deze ordenen is er tijdens het schrijven meer cognitieve capaciteit vrij om te focussen op het formuleren van die ideeën. Als na het schrijven van een eerste versie de tekst nog kritisch wordt doorgelezen en gereviseerd, wordt de tekst vaak nog beter (Fitzgerald & Markham, 1987). Peer-interactie tijdens en na het schrijven maakt leerlingen bewust van de communicatieve functie van schrijven en laat leerlingen ervaren hoe hun tekst overkomt op een lezer (Brakel Olson, 1990; Hoogeveen, 2013). Dit geeft belangrijke aanknopingspunten voor het reviseren van de tekst. Om een tekst nóg beter te maken is goede feedback (van peer of docent) onmisbaar (Holliway, 2004). Goede feedback geeft de leerling inzicht in zijn prestaties: wat doe ik al goed, wat kan beter? Naast het optimaliseren van het schrijfproces is kennis over verschillende teksttypen en tekstkenmerken on-ontbeerlijk om goede teksten te kunnen schrijven: een goed verhaal ziet er immers anders uit dan een instructietekst of een brief. Daarom is het belangrijk dat leerlingen via expliciete tekststructuurinstructie leren wat de kenmerken zijn van goede teksten in diverse genres (Crowhurst, 1991; Raphael & Kirschner, 1985). Met een combinatie van deze effectieve didactieken zouden leerlingen leerlingen in staat moeten zijn om goed te leren schrijven.

Toetsing en beoordeling van schrijfvaardigheid

Met alleen effectieve ingrediënten voor schrijflessen zijn we er natuurlijk nog niet. Om een echte verbeterslag te maken in het schrijfonderwijs is aandacht voor het toetsen en beoordelen van schrijven onontbeerlijk. Docenten moeten immers het niveau van hun leerlingen kunnen bepalen en hun schrijfontwikkeling over tijd kunnen volgen. Dit is echter niet makkelijk: in tegenstelling tot andere vakken en vakgebieden is het beoordelen van schrijven niet een kwestie van goed of fout. Het beoordelen van de kwaliteit van een tekst is sowieso al lastig, maar het vellen van een oordeel over de schrijfvaardigheid van een leerling wordt nog verder bemoeilijkt doordat (1) de schrijfprestatie van een leerling verschilt van taak tot taak en (2) beoordelaars onderling verschillen in hun oordeel over de kwaliteit van de tekst (o.a. Godshalk et al., 1966; Huang, 2009; Wesdorp, 1981). In hoofdstuk 3 en 4 staan respectievelijk de toetsing en beoordeling van schrijfvaardigheid centraal: hoe kun je op een valide en betrouwbare manier een oordeel vellen over de schrijfvaardigheid van een leerling?

In hoofdstuk 3 is onderzocht hoeveel taken en beoordelaars nodig zijn om een betrouwbaar oordeel over de individuele schrijfvaardigheid te kunnen vellen. Om hierachter te komen is in deze studie gebruik gemaakt van een steekproef van teksten geschreven door 67 leerlingen van groep 8 (Pullens, 2012). Deze leerlingen schreven twaalf teksten in vier verschillende genres en elke tekst werd gescoord door drie onafhankelijke beoordelaars. Deze beoordelaars waren zowel ervaren leraren als leraren in opleiding en ze beoordeelden de globale kwaliteit van de tekst met behulp van een gemiddelde ankertekst. Uit de resultaten bleek dat de verschillen tussen scores voor tekstkwaliteit maar voor 10% werden bepaald door de schrijfvaardigheid van leerlingen en voor 90% door factoren die niet (direct) gerelateerd waren aan individuele schrijfvaardigheid, zoals bijvoorbeeld het genre of onderwerp van de schrijftaak, of de beoordelaar. Hieruit bleek dat om te kunnen generaliseren naar de schrijfvaardigheid, leerlingen meerdere teksten in verschillende genres moeten schrijven waarbij elke tekst moet worden beoordeeld door ten minste twee beoordelaars.

In hoofdstuk 4 wordt nader ingegaan op de beoordelingsproblematiek bij schrijven. Als beoordelaars zonder enige sturing een holistisch oordeel moeten vellen over tekstkwaliteit, dat wil zeggen een oordeel over de tekst als geheel, blijken hun oordelen enorm uiteen te lopen (Diederich, French & Carlton, 1961). Om de betrouwbaarheid van de oordelen te verhogen wordt daarom vaak gebruik gemaakt van een analytische beoordelingsprocedure. Hierbij krijgen beoordelaars meer sturing in de manier waarop ze verschillende deelaspecten van de tekst moeten beoordelen. Deze analytische procedure lijkt echter wel ten koste te gaan van de validiteit van de oordelen: het is vaak moeilijk om op basis van de analytische scores nog iets te kunnen zeggen over de kwaliteit van de tekst in zijn geheel (Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). In dit onderzoek wordt daarom gekeken naar de betrouwbaarheid en validiteit van een nieuwe beoordelingsprocedure: een beoordelingsschaal met ankerteksten. De schaal bestaat uit vijf teksten die het spectrum van tekstkwaliteit van de doelgroep representeren en die oplopen in kwaliteit van heel slecht tot heel goed, zie bijlage I, J en K. Voor elke ankertekst is aangegeven waarom het een betere of slechtere tekst is dan de gemiddelde tekst. Beoordelaars moeten elke leerlingtekst vergelijken met de ankerteksten op de schaal om tot een holistisch oordeel voor tekstkwaliteit te komen. De assumptie is dat beoordelaars door de ankerteksten een beter beeld krijgen van de kwaliteit van een tekst als geheel en meer houvast hebben om tot een betrouwbare score te komen.

Om de kwaliteit van deze beoordelingsprocedure te onderzoeken hebben we de oordelen met beoordelingsschalen vergeleken met holistische en analytische oordelen. Uit de resultaten blijkt de beoordelingsprocedure met ankerschalen veelbelovend te zijn voor het beoordelen van tekstkwaliteit. Verschillen tussen beoordelaars zijn kleiner als ze de kwaliteit van teksten holistisch beoordelen met gebruik van ankerteksten dan wanneer ze holistisch beoordelen zonder ankerteksten. Daarnaast blijkt de toename van de betrouwbaarheid van oordelen met beoordelingsschalen niet ten koste te gaan van de validiteit van de oordelen, aangezien de tekst nog steeds als een geheel wordt beoordeeld en de oordelen met beoordelingsschalen beter generaliseerbaar zijn over taken dan het geval is bij analytische oordelen.

Wat betreft de praktische toepassing van schalen, blijkt het niet nodig dat voor elke schrijftaak een nieuwe schaal met ankerteksten wordt ontwikkeld. Als een beoordelaar teksten van een andere taak moet beoordelen dan de ankers op de schaal zijn de oordelen even betrouwbaar als wanneer de teksten van dezelfde taak zijn. Het lijkt wel raadzaam om voor elk genre een aparte schaal te ontwikkelen, aangezien de ankerteksten iets minder steun aan beoordelaars lijken te bieden bij de beoordeling van teksten uit een ander genre.

Vertaalslag van wetenschap naar praktijk: de lesmethode Tekster

De meta-analyse heeft concrete aanwijzingen opgeleverd hoe de inhoud van de instructie kan worden verbeterd, namelijk door het stellen van doelen, strategie-instructie, peer interactie, feedback en tekststructuurinstructie. Deze effectieve didactieken hebben we gecombineerd en geïntegreerd in het lesprogramma Tekster. De kern van Tekster is een algemene strategie voor de aanpak van schrijftaken, gebaseerd op de stappen van het schrijfproces: plannen, schrijven en reviseren. Om leerlingen te ondersteunen bij het aanleren van de strategie, gebruiken we acroniemen: VOS (Verzinnen, Ordenen, Schrijven) voor groep 6; DODO (Denken, Ordenen, Doen, Overlezen) voor groep 7; Ekster voor groep 8 (Eerst nadenken, Kiezen & ordenen, Schrijven, Teruglezen, Evalueren, Reviseren). Deze strategieën staan centraal in elke Teksterles.

Om een echte verbetering in het schrijfonderwijs te bewerkstelligen moet echter niet alleen de inhoud van instructie worden aangepakt, maar ook de manier waarop de instructie wordt gegeven. Onderzoek heeft laten zien dat de manier waarop instructie wordt gegeven net zo belangrijk is als de inhoud van de instructie (o.a. Hillocks, 1984). Om het leerproces van leerlingen optimaal te ondersteunen wordt bij Tekster gebruik gemaakt van verschillende vormen van instructie. De belangrijkste vorm van instructie in Tekster is observerend leren. Bij observerend leren kijkt een leerling naar een model (docent of peer) die hardop denkend een (deel van een) schrijftaak uitvoert (Fidalgo, Torrance, Rijlaarsdam, Van den Bergh, & Lourdes Alvarez, 2015). Het grote voordeel van observerend leren is dat leren en schrijven worden losgekoppeld, waardoor het leren alle aandacht krijgt (Rijlaarsdam, 2005). Andere vormen van instructie die worden toegepast in Tekster zijn expliciete instructie (O'Sullivan & Pressley, 1984), en scaffolding, oftewel ondersteuning tijdens het schrijven, die afneemt naarmate de leerling vaardiger wordt (Pearson & Gallagher, 1983).

In bijlage M staat een voorbeeldles van Tekster. Elke Teksterles start met expliciete instructie over het communicatieve doel dat centraal staat in de les en uitleg over de tekststructuur van het tekstype behorend bij dat communicatieve doel. Hierbij wordt gebruik gemaakt van voorbeelden van goede en slechte teksten ter illustratie of van modeling door de docent of peers. Na deze introductie krijgen leerlingen de schrijfopdracht. In elke les wordt het schrijfproces van de leerling ondersteund door middel van de stappen van de Teksterstrategie en door modeling door de docent. Daarnaast biedt de docent extra ondersteuning waar nodig door middel van scaffolding, bijvoorbeeld door verlengde instructie voor zwakkere schrijvers en feedback op maat gedurende het schrijfproces. Hoe docenten dit het beste kunnen aanpakken wordt beschreven in de docentenhandleiding die elke docent bij aanvang van het lesprogramma krijgt. In deze handleiding staat informatie over de achtergrond voor de opbouw van de lessen, modeling, feedback geven en het beoordelen van tekstkwaliteit. Daarnaast bevat de handleiding voor elke les een lesplan met hierin de tijdplanning van de les en per stap aanwijzingen instructie en organisatie van de les. Deze informatie komt ook aan bod in een voorbereidende training voor alle docenten die met het lesprogramma aan de slag gaan.

De effectiviteit van Tekster: Interventiestudie 1

In hoofdstuk 5 bespreken we de resultaten van de eerste interventiestudie naar de effecten van het lesprogramma Tekster op de schrijfprestaties van leerlingen. Het programma is getest bij 1420 leerlingen en 76 docenten van 60 klassen en 27 scholen. Voor deze studie maakten we gebruik van een switching replication design (Shadish, Cook, & Campbell, 2002), met twee groepen en drie meetmomenten. De ene helft van de docenten werkte met Tekster in de eerste periode van twee maanden, tussen het eerste en het tweede meetmoment. Zij gaven gedurende deze periode twee Teksterlessen per week aan de leerlingen van hun eigen klas. De tweede groep docenten volgde tijdens dezelfde periode het reguliere schrijfprogramma uit hun eigen taalmethode. Na het tweede meetmoment wisselden de groepen om en ging de tweede groep met Tekster aan de slag, terwijl de eerste groep weer hun reguliere schrijfprogramma uit de taalmethode ging volgen. Het voordeel van een switching replication design is dat binnen één studie het effect van de interventie in twee groepen wordt onderzocht, waardoor er bekeken kan worden of effecten repliceerbaar zijn. Ook is het door dit design mogelijk om de effecten op lange termijn te onderzoeken, aangezien het derde meetmoment voor de eerste groep een follow-up meting is. Daarnaast is het een ethisch design omdat alle deelnemende leerlingen uiteindelijk de interventie krijgen aangeboden.

Tijdens elk van de drie meetmomenten werd schrijfvaardigheid gemeten met drie schrijftaken in drie verschillende genres: verhalen, beschrijvingen en overtuigende brieven. In totaal schreef elke leerling dus 9 teksten, die elk beoordeeld werden door jury's van drie onafhankelijke beoordelaars. De resultaten zijn hierdoor te generaliseren over taken, genres en beoordelaars en geven dus een beter beeld van de schrijfvaardigheid van een leerling dan één leerlingtekst, beoordeeld door één beoordelaar.

De resultaten laten zien dat de interventie effect had op de schrijfprestaties van leerlingen. Na de interventie gingen leerlingen beter schrijven dan daarvoor, terwijl de prestaties van leerlingen in de controlegroep gelijk bleven. Deze vooruitgang was even groot voor beide groepen. Voor de leerlingen in de eerste groep was de vooruitgang na twee maanden nog steeds zichtbaar, wat erop lijkt te wijzen dat er bij leerlingen en/of docenten sprake is van een blijvende verandering. Uit logboeken en klasobservaties bleek dat docenten goed met het programma overweg konden: ze pasten de lessen toe zoals beschreven in de docentenhandleiding en maakten gebruik van de schrijfstrategie, het acroniem en modeling. Hierdoor hebben ze veel meer tijd aan het schrijfproces besteed dan dat ze normaal gesproken deden gedurende de schrijflessen.

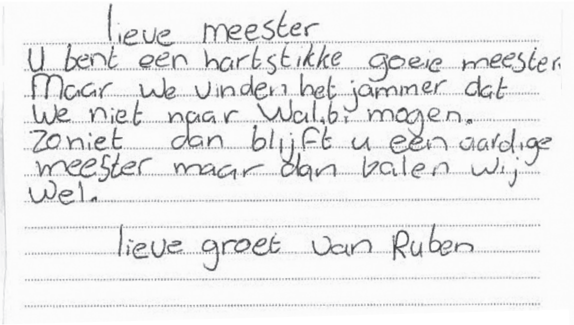
De resultaten lieten wel zien dat er grote verschillen waren tussen docenten in het toepassen van de belangrijkste componenten van het programma. Ook kwamen door organisatorische problemen niet alle docenten aan alle lessen toe, gemiddeld waren 10 van de 16 lessen gegeven. De verschillen tussen docenten bleken ook uit de leerlingresultaten: de effectiviteit van het programma verschilde van klas tot klas. Maar zelfs als er gecontroleerd werd voor deze verschillen tussen klassen, was het effect van de interventie zichtbaar. Generaliserend over klassen, maar ook over leerlingen en taken, is de sterkte van het effect 0.40, wat betekent dat leerlingen ongeveer een halve standaarddeviatie vooruitgaan. Als we dit vergelijken met de verschillen in schrijfprestatie tussen leerlingen in groep 6, 7 en 8 dan blijkt dat leerlingen na twee maanden onderwijs met Tekster ruim een half leerjaar vooruit zijn gegaan. Figuur 2 illustreert de verbetering in schrijfprestatie. In dit figuur zijn twee teksten van dezelfde leerling te zien, een voor en een na de interventie.

Figuur 2 Voorbeeld van overtuigende brieven geschreven door dezelfde leerling voor (links) en na (rechts) de interventie.

Schrijftaak: Schoolreisje naar pretpark

Jij wilt samen met je klasgenoten een dagje uit naar een pretpark. De meester of juf vindt dit niet zo'n goed idee. Jullie willen er toch alles aan doen om met z'n allen naar een pretpark te gaan.

Schrijf een briefje aan je meester of juf waarin je met goede argumenten probeert te overtuigen om toch met de klas naar dit pretpark te gaan. Maak in je brief duidelijk om wat voor pretpark het gaat.



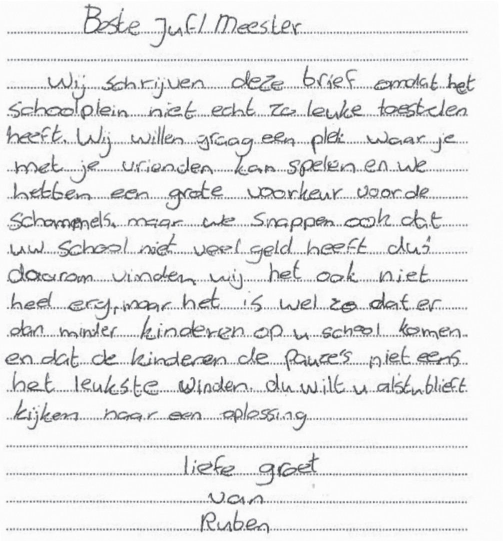
lieve meester
U bent een hartstikke goeie meester.
Maar we vinden het jammer dat
We niet naar Walibi mogen.
Zoniet dan blijft u een aardige
meester maar dan balen wij
Wel.

lieve groet van Ruben

Schrijftaak: Speeltoestel op het schoolplein

Tijdens de schoolpauze wil je graag leuk met je klasgenoten kunnen spelen, maar het schoolplein is heel saai. Jullie hebben daarom een geweldig idee: er moet een nieuw speeltoestel op het schoolplein komen!

Schrijf een briefje aan je meester of juf waarin je de school met goede argumenten probeert te overtuigen om een speeltoestel op het schoolplein te plaatsen. Maak in je brief duidelijk wat voor speeltoestel dit zou moeten zijn en waarom het belangrijk is dat dit speeltoestel er komt.



Beste Juf/Meester

Wij schrijven deze brief omdat het
schoolplein niet echt zo leuke toestelen
heeft. Wij willen graag een plek waar je
met je vrienden kan spelen en we
hebben een grote voorkeur voor de
Schommels. maar we Snappen ook dat
uw School niet veel geld heeft dus
daarom vinden wij het ook niet
heel erg, maar het is wel zo dat er
dan minder kinderen op u school komen
en dat de kinderen de pauze's niet eens
het leukste Winden. du wilt u alstublieft
kijken naar een oplossing

liefe groet
van
Ruben

Effectiviteit van de Teksterstrategie voor schrijven

Zoals hierboven is beschreven, is de kern van Tekster de schrijfstrategie VOS (groep 6), DODO (groep 7) en EKSTER (groep 8). Deze strategie is de rode draad door alle lessen; elke les wordt volgens de stappen van de strategie opgebouwd. Deze zijn gelijk aan de stappen van het schrijfproces, namelijk (1) een fase voor het schrijven waarin leerlingen ideeën genereren en organiseren (de prewriting fase), (2) een schrijffase waarin ze hun ideeën naar tekst vertalen en (3) een fase na het schrijven waarin ze hun teksten teruglezen, evalueren en reviseren (de postwriting fase). Het doel van de strategie is dat door de focus op slechts één onderdeel van het schrijfproces leerlingen minder cognitieve activiteiten tegelijkertijd hoeven uit te voeren. De assumptie was dat leerlingen hierdoor minder cognitieve overbelasting tijdens het schrijven zouden ervaren. Verder zou de strategie leerlingen ondersteunen bij het reguleren van hun schrijfproces, wat de kwaliteit van de tekst ten goede zou moeten komen. Deze aspecten hebben wij nader onderzocht in hoofdstuk 6 en 7.

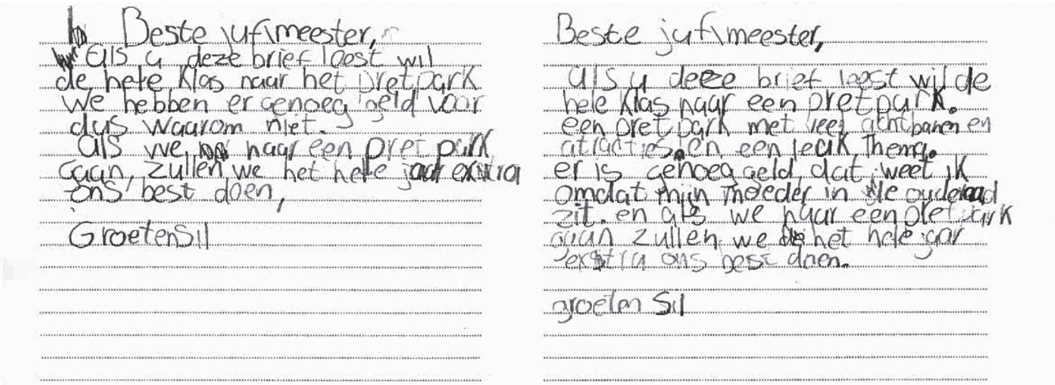
In hoofdstuk 6 is onderzocht of de kennis van leerlingen over schrijven en het schrijfproces is toegenomen door het aanleren van de schrijfstrategie en wat de invloed hiervan was op de kwaliteit van hun tekst. Leerlingen in de experimentele groep kregen drie Teksterlessen waarin de schrijfstrategie expliciet werd aangeleerd en waarin ze konden oefenen met het toepassen van de schrijfstrategie op verschillende soorten schrijftaken. Na de drie lessen moesten alle leerlingen zonder verdere expliciete instructie een brief schrijven aan een denkbeeldige klasgenoot waarin ze advies moesten geven over hoe je een goed cijfer voor schrijven kunt krijgen. Leerlingen in de controlegroep schreven dezelfde brief. Uit de gegeven adviezen bleek dat al na drie lessen de kennis van leerlingen over schrijven was toegenomen. Ze gaven niet alleen meer adviezen, ze gaven ook meer adviezen die betrekking hadden op hogere orde aspecten van schrijven, zoals het schrijfproces of de organisatie van een tekst, vergeleken met leerlingen uit de controlegroep. De adviezen van leerlingen die geen Teksterlessen hadden gekregen waren voornamelijk gericht op lagere orde aspecten zoals interpunctie, hoofdlettergebruik, spelling en grammatica. Ook al zijn 'adviezen' een indirecte meting van de kennis van een leerling (het betreft natuurlijk alleen de kennis die ze kunnen verbaliseren, en de adviezen die ze geven komen niet altijd tot uiting in hun eigen tekst), ze geven wel een goede indruk van de kennis die de leerlingen op dat moment hadden. Uit de resultaten bleek verder dat hoe meer leerlingen wisten van schrijven en het schrijfproces, hoe beter de kwaliteit van hun tekst was. Met andere woorden: door middel van het aanleren van een schrijfstrategie krijgen leerlingen meer kennis over schrijven, wat er vervolgens toe leidt dat ze betere teksten gaan schrijven.

In hoofdstuk 7 hebben we specifiek onderzocht wat de effectiviteit is van de stappen in de prewriting fase. Het uitgangspunt van deze stappen is dat als leerlingen al ideeën hebben bedacht en geordend voor het schrijven, ze zich tijdens het schrijven alleen maar bezig hoeven houden met het formuleren van woorden en zinnen, wat de kwaliteit van de tekst ten goede zou moeten komen.

We hebben dit onderzocht door bij de leerlingen die meegedaan hebben met de interventiestudie specifiek te analyseren hoe ze het kladpapier hebben gebruikt bij de schrijftaken op de drie verschillende meetmomenten. Doordat leerlingen niet werden geïnstrueerd om het kladpapier te gebruiken, geeft het gebruik ervan een goede indicatie of er een transfer is van de prewriting activiteiten die ze hebben geleerd in de lessen, namelijk het genereren van ideeën en deze vervolgens ordenen. Voor leerlingen in groep 6 werd gekeken naar het toepassen van prewriting bij het schrijven van verhalen, voor leerlingen in groep 7 werd gekeken naar het toepassen van prewriting bij het schrijven van beschrijvende teksten en voor groep 8 werd gekeken naar het toepassen van prewriting bij het schrijven van overtuigende brieven, om een goed beeld te krijgen van prewriting in de verschillende leerjaren en bij verschillende genres.

Zoals verwacht bleken leerlingen na de interventie meer aan prewriting te doen in vergelijking tot de leerlingen uit de controlegroep. Verder bleek prewriting samen te hangen met tekstkwaliteit: als leerlingen meer activiteiten voor het schrijven lieten zien, waren hun teksten beter. Ook hebben we gekeken naar verschillende vormen van prewriting. Waar leerlingen uit de controlegroep voornamelijk een kladversie van hun tekst schreven (als ze al aan prewriting deden), schreven leerlingen die Teksterlessen hebben gehad hun ideeën op en organiseerden deze voordat ze met het schrijven van de tekst begonnen. Deze strategieën komen overeen met wat de leerlingen bij Tekster hebben geleerd. Het type prewriting dat een leerling gebruikt blijkt effect te hebben op tekstkwaliteit. De meest effectieve prewriting strategie was het genereren en organiseren van ideeën voor het schrijven. Ook het maken van een lijst met ideeën in lijststructuur leidde tot betere teksten. Het minst effectief bleek het schrijven van een kladversie van de tekst te zijn. Figuur 3a (voor de interventie) en 3b (na de interventie) illustreren de verandering in het gebruik van prewriting en het effect op tekstkwaliteit. Deze voorbeelden laten duidelijk zien dat leerlingen teksten van een betere kwaliteit schrijven door eerst te plannen en daarna pas te gaan schrijven.

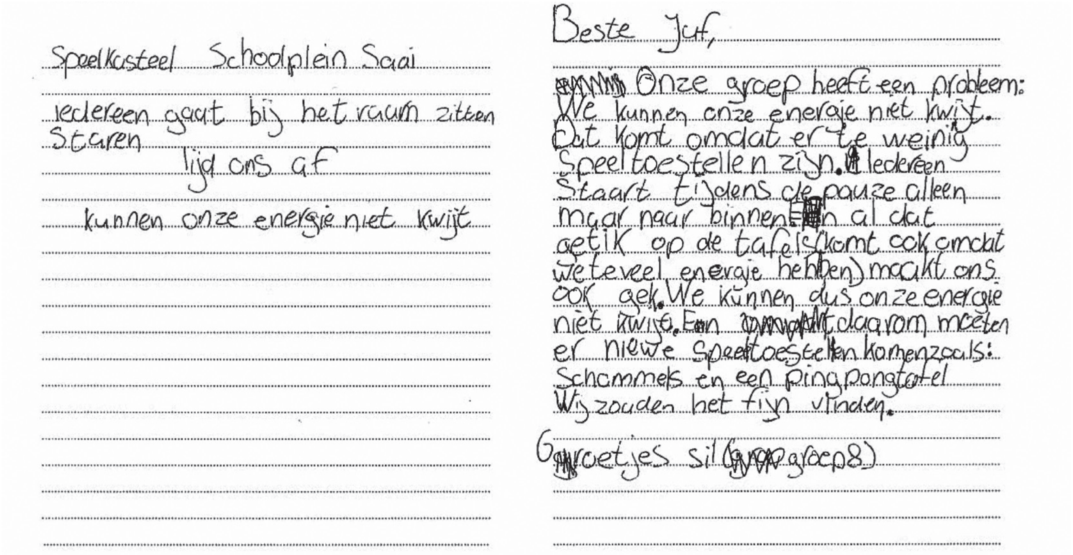
Figuur 3a **Een voorbeeld van prewriting (links) en de uiteindelijke tekst (rechts) geschreven door dezelfde leerling voor de interventie.**



Beste juf\meester,
als u deze brief loest wil
de hele Klas naar het pretpark
we hebben er genoeg geld voor
dus waarom niet.
als we naar een pret park
gaan, zullen we het hele jaar extra
ons best doen,
GroetenSil

Beste juf\meester,
als u deze brief leest wil de
hele Klas naar een pretpark.
een pretpark met veel achtbanen en
attracties. en een leuk Thema.
er is genoeg geld dat weet ik
omdat mijn moeder in de ouderaad
zit. en als we naar een pretpark
gaan zullen we het hele jaar
extra ons best doen.
groeten Sil

Figuur 3b Een voorbeeld van prewriting (links) en de uiteindelijke tekst (rechts) geschreven door dezelfde leerling na de interventie.



Speelkasteel Schoolplein Saai
Iedereen gaat bij het raam zitten
Staren
lijd ons af
kunnen onze energie niet kwijt

Beste Juf,
Onze groep heeft een probleem:
We kunnen onze energie niet kwijt.
Dat komt omdat er te weinig
Speeltoestellen zijn. Iedereen
Staart tijdens de pauze alleen
maar naar binnen. En al dat
getik op de tafels(komt ook omdat
we teveel energie hebben) maakt ons
ook gek. We kunnen dus onze energie
niet kwijt. Een daarom moeten
er niewe speeltoestellen komen zoals:
Schommels en een pingpongtafel
Wij zouden het fijn vinden.
Groetjes sil (groep8)

Het geven van feedback op teksten

Een andere belangrijke component van Tekster, naast de schrijfstrategie, is het geven van feedback. Uit de meta-analyse bleek dat feedback een heel effectief middel is om leerlingen beter te leren schrijven. Feedback verschaft leerlingen informatie over hoe hun tekst overkomt op een lezer. Verder geeft feedback de docent de mogelijkheid om instructie op maat te geven, aangepast aan de specifieke behoefte van de leerling. Een belangrijke voorwaarde voor het geven van effectieve feedback is dat de docent in staat moet zijn om op basis van de tekst in te kunnen schatten wat een leerling nodig heeft om vooruit te kunnen, en daar de feedback op te richten. Het alleen aanstippen van wat er allemaal beter kan in een tekst betekent immers niet dat een leerling hier ook mee uit de voeten kan (Biber, Nekrasova, & Horn, 2011). Om te onderzoeken of (en hoe) docenten hun feedback aanpassen aan de kwaliteit van de geschreven tekst is in hoofdstuk 8 de feedbackpraktijk van docenten onder de loep genomen. In totaal 14 basisschooldocenten van verschillende scholen gaven ieder feedback op 15 leerlingteksten in drie verschillende genres. De teksten waren verschillend in kwaliteit: er waren hele goede teksten, maar ook hele slechte, en alles daartussen in.

Uit de resultaten bleek dat docenten erg verschilden in de manier waarop ze feedback gaven. Over het algemeen gaven ze vooral feedback op dingen in de tekst die nog beter konden, en was evenveel van deze feedback gericht op problemen met hogere orde aspecten van de tekst (zoals inhoud, structuur, stijl) als op problemen met lagere orde aspecten (zoals interpunctie, spelling, grammatica, conventies). Opvallend was dat de kwaliteit van de tekst geen verschil maakte voor de inhoud van de feedback: goede en slechte teksten kregen evenveel feedback op lagere orde als op hogere orde aspecten. Wel gaven docenten gemiddeld iets meer positief commentaar op teksten van betere kwaliteit. Wat betreft de vorm bleek dat feedback voornamelijk directief van aard was. Docenten gaven eerder aan hoe de leerling problemen in de tekst kon oplossen, in plaats van het leerproces van de leerling te faciliteren door het stellen van vragen of het geven van lezersreacties. Een opmerkelijk resultaat was dat de mate van directieve feedback hoger was voor betere teksten. Verder gaven docenten, ongeacht de kwaliteit van de tekst evenveel commentaar in de tekst (tekstverbeteringen) als naast de tekst (algemeen commentaar).

Samenvattend valt op dat docenten zich bij het geven van de feedback voornamelijk richten op wat er beter kan in de tekst, in plaats van dat ze hun feedback proberen aan te sluiten op het niveau en de behoefte van de leerling. Extra ondersteuning voor docenten in het geven van effectieve feedback lijkt daarom raadzaam. Naar aanleiding van deze resultaten hebben wij het geven van effectieve feedback opgenomen als belangrijke component in het professionaliseringsprogramma dat we hebben ontwikkeld voor de tweede interventiestudie.

De effectiviteit van Tekster: Interventiestudie 2

Uit de resultaten van de eerste interventiestudie bleek dat er grote verschillen waren tussen klassen in de effectiviteit van het programma. Een mogelijke oorzaak hiervan zijn verschillen in de vaardigheid van docenten om schrijfonderwijs te geven. Dit was voor ons de aanleiding om bij de tweede interventiestudie, beschreven in hoofdstuk 9, meer aandacht te besteden aan scholing voor docenten. Evenals in de eerste interventiestudie maakten we in deze interventiestudie gebruik van een switching replication design en gaven de docenten de Teksterlessen aan hun eigen leerlingen. Aan deze studie deden in totaal 68 docenten en 1365 leerlingen mee van 25 verschillende scholen door heel Nederland.

Het grootste verschil met de eerste interventiestudie was de toevoeging van een professionali-seringscomponent voor docenten. De helft van de docenten ($N = 31$) volgde een trainingsprogramma gegeven door experts. Deze training bestond uit twee middagen, één voor de start van de interventie en één na 6 lessen. In de eerste training kregen ze uitleg over de belangrijke componenten van Tekster (de strategie en modeling) en over de opzet van het programma en de lessen. Tijdens deze bijeenkomst werd geoefend met modeling en werden gezamenlijk de eerste twee lessen voorbereid. De tweede trainingsbijeenkomst stond in het teken van het beoordelen van tekstkwaliteit en het geven van effectieve feedback, en was er de mogelijkheid om ervaringen uit te wisselen over de eerste lessen die gegeven waren. De docenten uit deze groep (trainers) trainden vervolgens hun collega's (trainees) die twee maanden later met het programma begonnen ($N = 37$). Hiermee konden we nagaan of de kennis en vaardigheden die docenten tijdens de trainingsbijeenkomsten leerden, overdraagbaar waren tussen collega's. We onderzochten of er verschillen waren tussen trainers en trainees in de effectiviteit van het programma wat betreft de schrijfprestaties van de leerlingen. Net als in de eerste interventiestudie werden de schrijfprestaties van de leerlingen gemeten op drie meetmomenten, met op elk meetmoment drie taken in drie verschillende genres. Daarnaast onderzochten we of de docenten door het lesprogramma en de training positiever werden over schrijven en schrijfonderwijs en of ze zich ook bekwamer voelden om schrijfonderwijs te geven.

Uit de resultaten bleek dat de schrijfprestaties van de leerlingen vooruitgingen door het volgen van het lesprogramma (ES = 0.55). We vonden geen verschillen tussen de prestaties van leerlingen van trainers of trainees. Gemiddeld gingen de schrijfprestaties van leerlingen na vier maanden Tekster meer dan anderhalf leerjaar vooruit, wat betekent dat leerlingen uit groep 6 die de lessen hadden gevolgd beter schreven dan leerlingen uit groep 7 die de lessen niet hadden gevolgd. Verder bleek uit vragenlijsten dat zowel trainers als trainees door het lesprogramma Tekster positiever werden over het geven van schrijfonderwijs en dat ze zich bekwamer voelden om zowel zwakke als sterke leerlingen te leren schrijven.

Het lesprogramma bleek dus zowel voor leerlingen als voor docenten positieve effecten te hebben. Er bleken geen verschillen te zijn tussen trainers en trainees wat erop lijkt te wijzen dat het programma goed overdraagbaar is tussen collega's binnen een school. Hoewel dit een zeer positief resultaat is, zegt het nog niets over de impact over het lesprogramma op het schrijfonderwijs in de praktijk. Hiervoor is meer informatie nodig over hoe docenten (componenten van) het programma in hun klas hebben toegepast en wat hun ervaringen hiermee zijn. Ook is het belangrijk om meer te weten te komen over de toegevoegde waarde van de training. Om hier meer inzicht in te krijgen, gebruikten we een mixed methods design waarin we de informatie van docenten uit logboeken, vragenlijsten, klasobservaties en interviews met elkaar combineerden. Uit deze mix van data bleek dat volgens docenten zowel de inhoud van hun lessen als hun manier van lesgeven waren veranderd door Tekster. De structuur van lessen zorgde ervoor dat ze leerlingen een strategie aanleerden voor het schrijven van teksten en meer aandacht hadden voor het schrijfproces. De inhoud van de lessen maakte dat ze meer gefocust waren op het communicatieve doel van de tekst. Docenten gaven aan dat ze tijdens hun instructie meer gebruik maakten van modeling en feedback dan dat ze daarvoor deden. Over deze componenten van Tekster waren de docenten erg positief.

Wel bleek uit de resultaten dat de docenten die de experttrainingen hadden gevolgd iets positiever waren over wat ze hadden geleerd dan de docenten die getraind waren door collega's. Dit was vooral het geval voor het toepassen van modeling, het beoordelen van de kwaliteit van leerlingteksten en het gebruiken van beoordelingsschalen voor beoordelen en feedback geven. Niet alle componenten van het lesprogramma blijken even gemakkelijk in het gebruik, waardoor een extra trainingsprogramma rondom Tekster een waardevolle toevoeging lijkt te zijn voor het verbeteren van het schrijfonderwijs. Dit pleit ook voor extra deskundigheidsbevordering over de inhoud en manier van instructie voor schrijven bij de lerarenopleidingen.

Tekster, het schrijfonderwijs van de toekomst

Ons onderzoek naar Tekster biedt veel aanknopingspunten voor de verbetering van het schrijfonderwijs. De resultaten laten zien dat meer ondersteuning bij het schrijven van teksten, meer aandacht voor schrijfproces en het communicatieve doel van een tekst helpen om de schrijfvaardigheid van leerlingen te verbeteren. Docenten gaven aan dat Tekster meer houvast bood bij hun schrijflessen dan de lessen uit de taalmethodes. Met name de schrijfstrategie speelde hierbij een grote rol: deze zorgde ervoor dat leerlingen beter voorbereid begonnen aan hun schrijftaak en dat ze meer ondersteuning hadden tijdens het schrijfproces. Doordat de schrijfopdrachten goed aansloten bij de belevingswereld en het niveau van de leerlingen, werden de schrijflessen ook leuker gevonden door zowel leerlingen als docenten. Verder bood Tekster instrumenten om schrijfprestaties van leerlingen in kaart te brengen. Hierdoor was het voor docenten makkelijker om gericht feedback te geven en om de schrijfontwikkeling van leerlingen te volgen.

Ook al zijn de bevindingen van ons onderzoek positief, het laat wel zien waar de knelpunten in het schrijfonderwijs zitten. Om echte verbetering in het schrijfonderwijs te bewerkstelligen is het nodig dat het onderwijs in schrijven hoog op de agenda komt van alle betrokkenen in het onderwijs: van schoolbesturen, schoolleiders en docenten. Het is opmerkelijk dat aan een basisvaardigheid als schrijven zo weinig tijd wordt besteed in vergelijking met bijvoorbeeld lezen en rekenen. Als we het belangrijk vinden dat leerlingen goede schrijvers worden, moeten we ook onderwijstijd en aandacht vrijmaken om ze dit te leren, door middel van gerichte instructie en ondersteuning tijdens het schrijven. Om leerlingen verder te helpen is het noodzakelijk dat hun schrijfprestaties systematisch gemonitord worden, zoals bij andere vakken al lang gebruikelijk is. De beoordelingsschalen die wij gebruikt hebben in ons onderzoek vormen hiertoe een eerste aanzet.

Daarnaast maakt dit onderzoek duidelijk hoe belangrijk de rol van de docent is bij het verbeteren van de schrijfvaardigheid van leerlingen, aangezien we grote verschillen zien tussen klassen in de effecten van de lesmethode. Zoals bij alle onderwijshervormingen is de docent in de klas de spil waar het allemaal om draait. Een echte verandering kunnen we alleen bewerkstelligen als we de docent beter toerusten om schrijfonderwijs te geven, door middel van scholing op het gebied van schrijfdidactiek en door de docent te voorzien van goed lesmateriaal. Een belangrijke rol bij het verbeteren van het schrijfonderwijs is dus weggelegd voor lerarenopleiders, onderwijsadviesdiensten en educatieve uitgeverijen.

Aanknopingspunten voor vervolgonderzoek

Onderzoek doen is keuzes maken. Hoewel wij in dit onderzoek een flink aantal aspecten hebben bekeken die een mogelijke bijdrage kunnen leveren aan het verbeteren van het schrijfonderwijs op de basisschool, zijn er nog een aantal aspecten die onderbelicht zijn gebleven of vragen om vervolgonderzoek. Zo hebben we wel uitgebreid onderzoek gedaan naar het effect van prewriting, een belangrijk onderdeel van de schrijfstrategie die centraal staat in Tekster, maar niet naar de postwriting activiteiten, zoals het evalueren en reviseren van teksten.

Verder hebben wij bij de instructie veel gebruik gemaakt van modeling, maar we hebben hiervan niet specifiek de effectiviteit onderzocht. Het is zeker zinvol om hier nader onderzoek naar te doen, ook om te bekijken hoe modeling zo optimaal mogelijk kan worden ingezet bij de schrijfinstructie. Werkt coping modeling beter dan mastery modeling, of is het juist andersom? Of verschilt dit voor goede of slechte schrijvers? Werkt modeling door de docent beter dan peer modeling, of juist niet?

Ons onderzoek laat verder zien dat er op het gebied van feedback geven nog een wereld te winnen valt. We hebben docenten getraind om feedback effectiever in te zetten, maar we hebben niet specifiek onderzocht hoe de training hun feedbackpraktijk heeft veranderd en wat de invloed van feedback is op de tekst van de leerling: worden teksten daadwerkelijk beter van meer gerichte feedback?

Door onze onderzoeksopzet hebben we informatie over de schrijfprestaties van leerlingen uit groep 6 tot en met 8, maar we hebben geen informatie hoe de schrijfvaardigheid van leerlingen zich tijdens deze jaren ontwikkelt. Hiervoor is longitudinaal onderzoek nodig, waarbij leerlingen van groep 6 tot en met 8 worden gevolgd.

Tot slot, wij hebben ons exclusief gericht op de bovenbouw van het basisonderwijs. Dit is echter geen eindpunt in de ontwikkeling van de leerling, maar het startpunt van een volgende fase: het voortgezet onderwijs, waar schrijven een belangrijk middel is om te leren bij andere vakken dan alleen Nederlands. De strategiegerichte aanpak zou ook hiervoor geschikt kunnen zijn, maar naar hoe dit succesvol zou kunnen worden geïmplementeerd is eveneens meer onderzoek nodig.

Algemene conclusie

Uit dit proefschrift blijkt dat de lesmethode Tekster, waarin een combinatie van effectieve didactieken wordt aangeboden, een bijdrage kan leveren aan de verbetering van het schrijfonderwijs in de bovenbouw van het basisonderwijs. Tekster is ontwikkeld met en voor docenten, wat heeft bijgedragen aan de effectiviteit van de methode en het gemak waarmee het te gebruiken is in het klaslokaal. Leerlingen gaan enorm vooruit in schrijven en leren met Tekster een strategie die ze ook kunnen inzetten bij het schrijven in andere vakken. In combinatie met een training voor docenten en instrumenten voor toetsing en beoordeling biedt Tekster docenten een volledig pakket waarmee zij hun leerlingen een goede basis kunnen geven voor hun verdere school- en beroepscarrière. Zowel leerlingen als docenten zijn erg enthousiast over het werken met Tekster: “Door Tekster heb ik een 10 voor schrijven op mijn rapport” (Tom, leerling groep 8) en “Tekster is een recept dat altijd tot een lekkere taart leidt” (Joost, docent groep 8). Kortom, met Tekster is leren schrijven een feestje!

Samenvatting

DE NAMEN DIE HET VERSCHIL HEBBEN GEMAAKT

Huub van den Bergh, onze held & godfather & inspirator & grootste fan & de H van HuMoR.

Ted Sanders & Anton Béguin, de wijze heren en (co-) promotoren: dank voor het kritisch meelezen en meedenken.

Gert Rijlaarsdam, onze onvermoeibare schrijfonderzoekexpert & wetenschappelijk supporter.

De bevlogen docenten uit de ontwikkelteams die samen met ons de Teksterlessen hebben gemaakt: Nellie, Tineke, Edith, Leanne, Esther, Mariëlle, Jeroen, Riesje, Eric, Carlijn, Marga, Melline, Mariëlle, Johanna; en Rob voor het monteren van de filmpjes.
Alle 144 leerkrachten en 2766 leerlingen van de 52 basisscholen* die van ons Teksteronderzoek een waar feestje hebben gemaakt! Door jullie ligt hier een beestachtig prachtig proefschrift. Bijzonder dank aan Liesbeth Mol, zonder jou geen Tekster the movie; Ornette Troost, Els de Groot, Jikkie Bosgraaf & Yunus Meijerink, Teksterfans van het eerste uur en onze eerste echte Teksterklanten die dit avontuur met ons gaan voortzetten; Petra Petersen voor alle enthousiaste retweets op Twitter.

Onze kopieerkoning Don van den Bergh en stickerkoningin Marlène Ottevanger, zonder wie de 24894 leerlingteksten nooit in drievoud waren gekopieerd en geanonimiseerd. De gouden kracht Eline ‘heb je nog een klusje’ Post. Eva Noortman voor het invoeren van alle lessen. Onze inpakkers: Rogier & Mattis & Kees. Paul, Joop en Leo voor het twee trappen op sjouwen van alle dozen met ingevulde leerlingboekjes. De 65 leerkrachten die maar liefst 74682 teksten hebben beoordeeld en onvermoeibaar bleven genieten van het lezen van alle pareltjes van teksten.

Rémi Bouwer, onze Teksterdesigner, voor het “piemelen” en gouden eieren leggen.

UtrechtIncers & onze mentoren, Ferry van Halem en Steven van den Tol, voor het meehelpen uitvliegen van Tekster. Frans Camp, voor het financiële plaatje.

Onze hulpelven: Marrit, Fien, Renee & Eva.

De liefjes Ron en Elske ♥, want zonder liefde geen proefschrift.

* Alle 52 basisscholen die met Tekster hebben gewerkt:

Schooljaar 2013 -2014

Al-Ummah, Enschede

Bosch en Hovenschool, Heemstede

C.b.s. de Weele, Strijen

De Heliotroop, Almere

De Kosmos, Hoogland

De Lage Weijkens, Loon op Zand

De Odyssee, Amsterdam

De Spil, Klazienaveen

De Springplank, Meerkerk

De Stappen, Wouw

De Talenten, Haarlem

De Twee Wieken, Zwijndrecht

De Wegwijzer, Alkmaar

De Wegwijzer, Spakenburg

De Wegwijzer, Teteringen

Eben-Haezer, Drachten

Freinetschool, Delft

Heilig Hartschool, Mussel (Kopstukken)

Het Talent, Roosendaal

Lodijke, Bergen op Zoom

Maria, Oirlo

Prins Bernhard, Tilburg

Rehobôthschool, Barendrecht

RK Daltonschool De Driesprong, Soest

Sint Jozef, Amsterdam

St. Pancratiusschool, Brummen

Violenschool, Hilversum

Schooljaar 2014-2015

Albertine Agnesschool, Leeuwarden

De Brug, Utrecht

De Fontein, Alkmaar

De Grunslach, Oudega (Smallingerland)

De Kosmos, Hoogland

De Lage Weijkens, Loon op Zand

De Meent, Waalre

De Odyssee, Amsterdam

De Stappen, Wouw

De Vlindervallei, Amersfoort

De Wegwijzer, Spakenburg

Dr. H. Bavinckschool, Haarlem

Eben-Haezer, Drachten

Gertrudisschool, Utrecht

Goeman Borgesiusschool, Amsterdam

Groen van Prinstererschool, Dedemsvaart

Isandraschool, IJzendoorn

Jenaplan basisschool 't Hoge Land, Epe

JF Kennedyschool, Den Haag

Johan Willem Frisoschool, Leeuwarden

Marcusschool, Utrecht

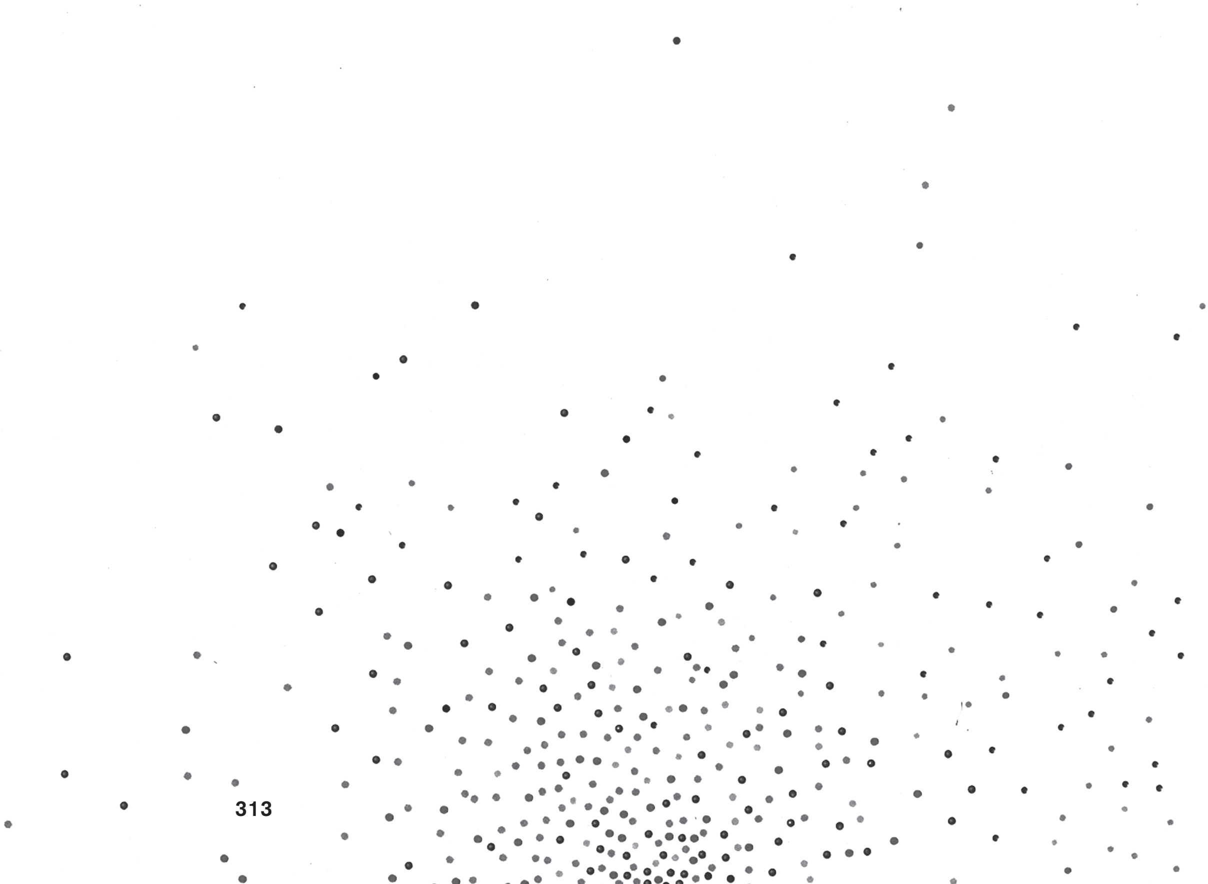
Maria, Oirlo

Tienercollege, Gorinchem

Titus Brandsma school, Hengelo

Violenschool, Hilversum

En dan zijn er nog zoveel mensen die ons langs de zijlijn zo enorm hebben aangemoedigd. Dankzij jullie steun kregen wij en Tekster echt vleugels! Ook jullie willen we graag bedanken, kom bij ons langs voor een unieke persoonlijke boodschap/opdracht in jouw eigen exemplaar, op 2 september bij de verdediging, of bij een latere gelegenheid.



CURRICULUM VITAE

Renske Bouwer was born on April 20th 1984 in Purmerend, the Netherlands. After completing secondary education in 2002, she started to study Psychology at the Radboud University (RU) in Nijmegen. In 2006 she obtained a BA degree in Social Psychology. She then got selected for a research master in Behavioural Science at the same university, which she completed in 2008 (bene meritum). In the three following years (2009-2012) she worked as a student tutor and lecturer for a variety of BA Psychology courses, amongst which academic writing. During this time she was certified with a Basic Teaching Qualification and got the opportunity to develop the examination policy of the BA Psychology. She started in 2012 with her PhD research on improving elementary students' writing skills at the Utrecht Institute of Linguistics OTS, in collaboration with the Dutch Institute for Educational Measurement (Cito) and the teacher training department of the Avans University of Applied Sciences (Pabo Avans Hogeschool). Her focus is in particular on how students' writing proficiency can be assessed in a valid and reliable way, and how teachers can use this information to provide effective feedback to their students.

Monica Koster was born on May 29th 1967 in Den Helder, the Netherlands. She completed her secondary education in 1986, and after obtaining teaching qualifications in both English and Dutch in 1991, she got a Master's degree in Dutch Language and Literature in 1995 at Utrecht University. For several years she has worked as a special educational needs teacher at an elementary school, combined with her own practice. Through the years she worked with elementary and secondary students with various learning and developmental problems, such as dyslexia and autism. In 2010 she got selected for a research master in Educational Sciences at the University of Amsterdam, which she completed in 2013. By then, she had already started with her PhD project on improving the writing skills of students in the upper elementary students at the Utrecht Institute of Linguistics OTS of Utrecht University, in collaboration with the Dutch Institute for Educational Measurement (Cito) and the teacher training department of the Avans University of Applied Sciences (Pabo Avans Hogeschool). Within the project she focused in particular on establishing effective didactical approaches to improve the teaching of writing in the upper grades of elementary education. In close collaboration with elementary teachers, she developed the lessons for the teaching program Tekster.

Foto's Marloes Herijgers

During their joint collaboration, Monica and Renske became part of a large national and international network of writing researchers. They initiated a PhD group on writing research, which now counts more than 20 members in the Netherlands and Flanders. In 2014, in collaboration with researchers of the University of Amsterdam, they organized the international EARLI SIG Conference on Writing Research in Amsterdam, which was preceded by a two-day research school for PhD students at the Utrecht University. In 2015 they received a Best of JURE paper award for their joint paper on the effectiveness of Tekster (see chapter 5 in this dissertation), and an honourable mention at the NRO-VOR praktijkprijs [award to promote the implementation of research into practice] for the lesson program Tekster.

Since 2015 they have worked on launching the evidence-based writing program Tekster into the market, supported by the university business incubator UtrechtInc (6th place on the UBI index), where they were selected for the Science Venture program in november 2015. Tekster is now for sale, and as of the school year 2016-2017 several elementary schools will start working with the program. Future plans include a digital version of Tekster and expanding the program to the first grades of secondary education. For more information see www.tekster.nl or follow us on @TeksterNL.

ICO DISSERTATION SERIES

In the ICO Dissertation Series dissertations are published of graduate students from faculties and institutes on educational research within the ICO Partner Universities: Eindhoven University of Technology, Leiden University, Maastricht University, Open University of the Netherlands, University of Amsterdam, University of Twente, Utrecht University, VU University Amsterdam, and Wageningen University, and formerly University of Groningen (until 2006), Radboud University Nijmegen (until 2004), and Tilburg University (until 2002). The University of Groningen, University of Antwerp, University of Ghent, and the Erasmus University Rotterdam have been 'ICO 'Network partner' in 2010 and 2011. From 2012 onwards, these ICO Network partners are full ICO partners, and from that period their dissertations will be added to this dissertation series.

286. Karimi, S (14-3-2014) *Analysing and Promoting Entrepreneurship in Iranian Higher Education: Entrepreneurial Attitudes, Intentions and Opportunity Identification.* Wageningen: Wageningen University

287. Frambach, J.M. (26-3-2014) *The Cultural Complexity of problem-based learning across the world.* Maastricht: Maastricht University.

288. Mascareno, M.N. (11-4-2014) *Learning Opportunities in Kindergarten Classrooms. Teacher-child interactions and child developmental outcomes.* Groningen: University of Groningen

289. Bakker, M. (16-04-2014) *Using mini-games for learning multiplication and division: A longitudinal effect study.* Utrecht: Utrecht University

290. Loon, Marriette van (8-5-2014) *Fostering Monitoring and Regulation of Learning.* Maastricht: Maastricht University

291. Coninx, N.S. (28-05-2014) *Measuring effectiveness of synchronous coaching using bug-in-ear device of pre-service teachers.* Eindhoven: Eindhoven University of Technology.

292. Baars, M.A. (6-6-2014) *Instructional Strategies for Improving Self-Monitoring of Learning to Solve Problems.* Rotterdam: Erasmus University Rotterdam.

293. Hu, Y. (26-6-2014) *The role of research in university teaching: A comparison of Chinese and Dutch teachers.* Leiden: Leiden university

294. Rutten, N.P.G. (5-9-2014) *Teaching with simulations.* Enschede: University of Twente

295. Rijt, J.W.H. van der, (11-9-2014) *Instilling a thirst for learning. Understanding the role of proactive feedback and help seeking in stimulating workplace learning.* Maastricht: Maastricht University

296. Engelen, J. (11-09-2014) *Comprehending Texts and Pictures: Interactions Between Linguistic and Visual Processes in Children and Adults*. Rotterdam: Erasmus University Rotterdam

297. Khaled, A.E. (7-10-2014) *Innovations in Hands-on Simulations for Competence Development. Authenticity and ownership of learning and their effects on student learning in secondary and higher vocational education*. Wageningen: Wageningen University

298. Gaikhorst, L. (29-10-2014) *Supporting beginning teachers in urban environments.* Amsterdam: University of Amsterdam

299. Wijnia, L. (14-11-2014) *Motivation and Achievement in Problem-Based Learning: The Role of Interest, Tutors, and Self-Directed Study*. Rotterdam: Erasmus University Rotterdam

300. Gabelica, C. (4-12-2014) *Moving Teams Forward. Effects of feedback and team reflexivity on team performance*. Maastricht: Maastricht University.

301. Leenaars, F.A.J. (10-12-2014) *Drawing gears and chains of reasoning. Enschede:* University of Twente

302. Huizinga, T. (12-12-2014) *Developing curriculum design expertise through teacher design teams*. Enschede: University of Twente

303. Strien, J.L.H. van (19-12-2014) *Who to Trust and What to Believe? Effects of Prior Attitudes and Epistemic Beliefs on Processing and Justification of Conflicting Information From Multiple Sources*. Heerlen: Open University of the Netherlands.

304. Wijaya, A. (21-01-2015) *Context-based mathematics tasks in Indonesia: Towards better practice and achievemen*t Utrecht: Utrecht University

305. Goossens, N.A.M.C. (22-01-2015) *Distributed Practice and Retrieval Practice in Primary School Vocabulary Learning* Rotterdam: Erasmus University

306. Jupri, A. (28-01-2015) *The use of applets to improve Indonesian student performance in algebra*. Utrecht: Utrecht University

307. Griethuijsen, R.A.L.F. van (11-03-2015) *Relationships between students' interest in science, views of science and science teaching in upper primary and lower secondary education* Eindhoven: Eindhoven University of Technology

308. De Smet, C. (11-05-2015) *Using a learning management system in secondary education: Design and implementation characteristics of learning paths* Ghent: Ghent University

309. Aesaert, K. (19-05-2015) *Identification and assessment of digital competences in primary education G* Ghent: Ghent University

310. Ardies, J. (22-05-2015) *Students' attitudes towards technology. A cross-sectional and longitudinal study in secondary education* Antwerp: University of Antwerp

311. Donker, A.S. (11-06-2015) *Towards effective learning strategies* Groningen: University of Groningen

312. Veldhuis, M. (24-06-2015) *Improving classroom assessment in primary mathematics education* Utrecht: Utrecht University

313. Leeuwen, A. van (30-06-2015) *Teacher Regulation of CSCL: Exploring the complexity of teacher regulation and the supporting role of learning analytics* Utrecht: Utrecht University

314. Boschman, F.B. (28-08-2015) *Collaborative design of ICT-rich early literacy learning material: Design talk in teacher teams* Enschede: Twente University.

315. Dijk, M.L. van (04-09-2015)*Physical Activity, Cognitive Performance and Academic Achievement in Adolescents* Heerlen: Open University of the Netherlands

316. Jaarsma, T. (04-09-2015)*Expertise Development Under the Microscope: Visual Problem Solving in Clinical Pathology.* Heerlen: Open University of the Netherlands

317. Isac, M.M. (01-10-2015) *Effective civic and citizenship education. A cross-cultural perspective* Groningen: University of Groningen.

318. Spelt, E.J.H. (26-10-2015) *Teaching and learning of interdisciplinary thinking in higher education in engineering.* Wageningen: Wageningen University.

319. Ritzema, E.S. (05-11-2015) *Professional development in data use: The effects of primary school teacher training on teaching practices and students' mathematical proficiency* Groningen: University of Groningen

320. Gijselaers H.J.M. (06-11-2015) *Biological Lifestyle Factors in Adult Distance Education: Predicting Cognitive and Learning Performance.* Heerlen: Open University of the Netherlands

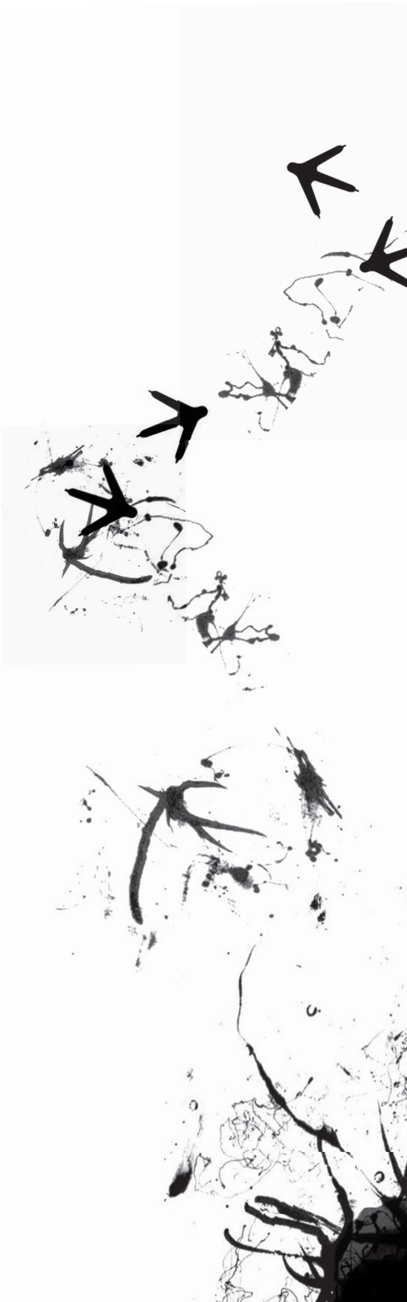
321. Oude Groote Beverborg, A. (12-11-2015). *Fostering sustained teacher learning: Co-creating purposeful and empowering workplaces* Enschede: Twente University.

322. Dijkstra, E.M. *(13-11-2015) Teaching High-Ability Pupils in Early Primary School.* Heerlen: Open University of the Netherlands

323. Want, A.C. van der (17-11-2015) *Teachers' Interpersonal Role Identity Eindhoven:* Eindhoven University of Technology

324. Wal, M.M van der (4-12-2015) *The Role of Computer Models in Social Learning for Participatory Natural Resource Management* Heerlen: Open University of the Netherlands

325. Herten, M. van (11-12-2015) *Learning communities, informal learning and the humanities. An empirical study of book discussion groups.* Heerlen: Open University of the Netherlands



www.tekster.nl

TEKSTER