

Mapping synthesis writing in various levels of Dutch upper-secondary education

A national baseline study on text quality, writing process and students' perspectives on writing

N. Vandermeulen, S. De Maeyer, E. Van Steendam, M. Lesterhuis, H. van den Bergh, en G. Rijlaarsdam

Abstract

Writing a synthesis text - a text that integrates information from different sources - forms part of the educational curriculum of Dutch secondary education. Representative information on students' synthesis writing skills is currently missing. Therefore, we carried out a national survey on synthesis writing in the three highest grades of pre-university education in the Netherlands. The aim was to map synthesis writing on three aspects: text quality, writing process and students' perspectives on writing. A large and representative sample of 658 students participated. Each participant wrote several synthesis texts. Text quality was rated with benchmark texts; writing processes were registered with keystroke logging software and a questionnaire measured students' perspectives on writing. Multilevel analyses were used to identify the effect of grade, gender and genre (argumentative/ informative synthesis) on text quality and writing process, and the effect of grade and gender on perspectives. This national survey is a descriptive study providing information on the current state of synthesis writing of Dutch students: how well do students perform on synthesis tasks?, how do they write synthesis tasks?, and what are their perspectives on synthesis writing? Moreover, this study serves as a baseline for future research.

Keywords: baseline study, synthesis writing, keystroke logging, writing process, writing education

1 Introduction

1.1 Synthesis writing

Source-based writing

Writing synthesis texts -texts which integrate information from different sources- is challenging, given the cognitively demanding nature of this task (Martínez et al., 2015; Mateos et al., 2008; Solé et al., 2013). The process of source-based writing, such as synthesis writing, involves both reading and writing, which led Spivey and King (1989) to label it as a hybrid task. The complexity of synthesis writing does not call for a simple "reading-then-writing" strategy. Rather, it involves a complex interplay of reading and writing sub-processes. During the writing process, students alternate between reader and writer roles as they read sources, select relevant information from the sources, compare and contrast the information from the different source texts to each other, write and revise the actual text. Key to synthesis writing is the integration process which encompasses connecting the ideas from the different source texts by organising and structuring them around a central theme in the source-independent target text (Solé et al., 2013; Spivey & King, 1989).

The term *synthesis task* is used for a rather wide range of source-based tasks. What all synthesis tasks have in common is that they require the integration of relevant information from sources. The diversity of synthesis tasks is reflected in previous research. An important distinction is the communicative function: the argumentative synthesis genre (Anmarkrud et al., 2014; Mateos et al., 2008; Solé et al., 2013), or the informative genre (Boscolo et al., 2007). Also the number of sources and the relation between the sources vary in synthesis studies. Boscolo et al. (2007) and Spivey &

King (1989) used tasks based on three sources, participants in the study of Anmarkrud et al. (2014) received six sources. Some studies choose to focus on conflicting sources (Anmarkrud et al., 2014; Du & List, 2020), while others provided complementary sources (Spivey & King, 1989). All these varying textual features may have an impact on the integration process (Barzilai et al., 2018). In the present study, we will vary some of the features of the synthesis task systematically to assure the generalisability of the findings.

Dutch educational curriculum

In the Netherlands, expert groups have designed a national frame of reference for Dutch language education, including writing. This framework contains several goals and formulates what a student should master at the end of a certain educational level. The curriculum postulates that upper-secondary students should be able to “synthesise information from various sources into one text” and to “write a text [...] on complex themes in which they stress relevant information, based on various sources” (Expertgroep Doorlopende Leerlijnen, 2009, p. 15). It is important that upper-secondary students develop their synthesis writing proficiency as, in higher education, they will need skills like selecting relevant information from sources, and integrating this information into a new and source-independent text (Feddemma & Hoek, 2018). However, as Van Ockenburg, Van Weijen and Rijlaarsdam (2018) point out, students implicitly practice such writing when writing an essay, but generally it is not a writing activity that is explicitly taught in Dutch schools, not in literacy lessons, nor in other school subjects.

1.2 Baseline studies

Importance

National surveys provide important information on the current state and the progress in several educational disciplines (De Glopper, 1988; National Center for Education Statistics, 2012). National surveys result in representative information on what students can accomplish in a certain grade.

Moreover, it allows to map the development of skills over the grades. In this way, national surveys evaluate the state of affairs and the progress in a certain educational field. The obtained information can be used to adapt the curriculum or to decide on areas of focus to further shape education. Moreover, national studies can serve as a baseline for other studies.

National surveys in the Netherlands

Cito is a national educational measurement organisation that carries out national assessment studies in several educational domains in the Netherlands. Currently, no national study on the writing skills of secondary students is available. There have been, however, several studies on the writing skills of pupils in Dutch primary education (Sijtsma, 1997; Sijtsma et al., 1998; Zwarts, 1990) and a feasibility study for a national assessment in secondary education (Kuhlemeier & Van den Bergh, 1990). The most recent report on the writing skills of Dutch primary students dates from 2010 (Inspectie van het Onderwijs, 2010). Results of this national study carried out in 2009 indicated that there is a significant difference in writing skills (measured as text quality) between the different grades in primary education, with the higher grades scoring higher than the lower grades. However, the report also concluded that there is a great discrepancy between the writing skills of the Dutch pupils and the goals as postulated in the educational curriculum framework. Following up on this national study, Kuhlemeier, Van Til, and Van den Bergh (2014) pointed out that schools tend not to prioritise writing education, and that, within writing education, there is little attention to the development of writing skills.

Grade, gender, genre

When collecting representative data for a national survey on students' writing skills, we do not only obtain information on students' individual writing skills, but also on the relation between those writing skills and student factors and task factors that could explain variation in writing skills. In this

study we chose to describe Dutch students' synthesis writing skill while taking into account two student factors, namely grade and gender, and one task factor, namely genre. First, we assess the grade effect, following previous studies that showed that writing skills evolve over the schooling years (Drijbooms, 2016; Mateos & Solé, 2009). A second factor to be included in our study was gender as previous research (see Cordeiro, Castro, & Limpo (2018) for an overview) has shown that girls tend to outperform boys when it comes to a variety of writing skills in all grades, also writing conceptions are affected by gender (Villalón et al., 2015). And thirdly, we assess whether synthesis writing skill is generalisable across genres, as studies showed that writing performance may depend on communicative function or genre (Bouwer et al., 2015), and the definition of synthesis tasks encompasses both informative and argumentative functions.

1.3 Writing product, process and perspectives

This study aims to provide a national baseline on synthesis writing for the upper grades of pre-university education. To provide a fairly complete view on synthesis writing, the study will focus on three aspects of writing, namely the quality of the product, the writing process, and students' perspectives on writing. The first indicator of writing skills is the quality of the written texts. Text quality gives information on *how well* students perform. Writing skills tend to develop over the grades as text quality increases in higher grade students; this is also the case for synthesis texts, though previous research indicates that the proportion of successful synthesis texts is low, even for university students (Mateos & Solé, 2009).

A second important aspect of writing skill is the writing process. Studying the writing process will provide us with an insight into *how* students write a synthesis text. The temporal distribution of cognitive activities in the process can predict (part of) the quality of the text (Van den Bergh & Rijlaarsdam, 2001). Studies by Martínez et al. (2015) and Mateos and Solé (2009) show that higher-grade, and thus more experienced, students

tend to adopt a less linear writing approach when writing a synthesis text. This involves a more recursive process in which reading and writing activities alternate and recur throughout the process.

A third factor under study are students' perspectives on writing. For this study, we will include several perspectives on *affective and cognitive aspects* of writing. These aspects relate to students' writing skills and may change over time (Graham, 2018).

2 Aim of the present study

In this study, we report on a national survey on synthesis writing carried out in the three grades of upper-secondary education in the Netherlands. As a national survey study, this study is purely descriptive. Three aspects of the students' writing are reported: the students' writing performance based on the quality of their written texts, the students' writing processes, and their perspectives on writing. The aim of this study is three-fold as we analyse the effect of grade, gender, and genre on students' synthesis writing. We also explore possible interactions of grade with gender and genre. We will address the following three research questions:

- a) What is the effect of grade on (1) writing performance, (2) writing process, and (3) perspectives on writing?
- b) What is the effect of gender on (1) writing performance, (2) writing process, and (3) perspectives on writing? And does the effect of grade differ for gender?
- c) What is the effect of genre on (1) writing performance, and (2) writing process? And does the effect of grade differ for the two genres?

The present study thus aims to describe the development of text quality, writing process and perspectives over the three highest grades of secondary education, and how this differs for argumentative and informative synthesis texts, and for boys and girls. We will offer a fairly complete view on the current state of synthesis writing and a baseline for future (intervention) research.

3 Method

3.1 Sampling procedure

Sample size

The goal of this study calls for a sample that is representative for the population of Dutch children in the last three grades of upper secondary education (grades 10-11-12). Deciding on a proper sampling for such a national survey study is a challenge. Sample simulations taking into account cluster effects (between school variance) were used to decide on the sample size. Table 1 shows that the standard error of a sample decreases if the proportion of variance between schools (intraclass correlation) increases. For instance, if the proportions of variance between schools equals 5%, and we sample 4 students per grade, then the standard error of the mean is approximated as .07. This indicates that a 95% confidence interval for the mean ranges from ($z_{95\%} \cdot .07 =$) - .14 *SD* to .14 *SD*. The precision increases slightly if the number of students increases, and a little more if the number of schools increases. Based on these simulations, we have chosen to sample 40 schools and 8 students per grade.

Sampling frame

The basis for constructing our sampling frame consisted of a data sheet with the number of students enrolled in pre-university education, clustered in 486 schools. The data were obtained via *DUO (Dienst Uitvoering Onderwijs*, the Education Office of the Dutch Ministry). We used the most up-to-date datasheet available at the moment of data collection (February 2016), that is, the enrolment data of school year 2014-2015.

Sampling method

To obtain a representative sample of pre-university students, a two-stage cluster sampling method was used. In the first stage, 40 schools were selected proportionally to their size; in the second stage, 24 students (8 for each grade) within these schools were selected.

The schools (i.e., the first-stage clusters) were selected by a systematic protocol. To make the sample as representative of the population as possible, schools were sampled proportionally to size. That is, schools with a higher number of students had a higher chance to be selected than schools with a lower number of students. For sample size *n* of 40 schools, we divided the population of 42 253 grade 10 students by 40 (42253/40= 1056.33). Starting at a random school, we then selected the schools containing the $n_{1,2,3...40} \cdot 1056$ pupil. Following these steps we obtained a sample frame of 40 schools that were invited to participate in the national baseline study.

Anticipating a low response, we performed the sample procedure twice more to create two backup sample frames. So, in the case a school from the first sample did not want to participate, a school from the second (and later third) sample was contacted. From the 3 sample frames, we found 36 schools willing to participate (10 schools from the main sample, 11 schools from backup sample 1 and 15 schools from backup sample 2). Per sample frame, the response rate was an acceptable 25% or higher. Apart from the 36 schools selected via systematic sampling, six more schools that expressed their interest to participate were included in our sample. So, in total 43 school agreed to participate.

Table 1
Expected standard error of the mean (SD) for different numbers of students per grade ($N_{students}$) and different numbers of schools for three values of intraclass correlation (.05, .10, and .20) and four writing tasks per student (estimates based on 5000 samples each)

$N_{students}$	$N_{schools}$								
	30			40			50		
	.05	.10	.20	.05	.10	.20	.05	.10	.20
4	.07	.07	.09	.05	.06	.08	.04	.05	.07
8	.06	.07	.09	.04	.06	.08	.04	.05	.07
12	.05	.06	.08	.04	.06	.08	.04	.05	.07

The second-stage sample of participants was selected by a simple random sampling protocol within each first stage sample cluster. Per school and per grade, students were selected randomly. We aimed at 8 participating students per school per grade. Anticipating participant drop-out, 10 students per grade per school were selected. On average, 8.02 students participated per school per grade ($SD= 2.27$).

3.2 Participants

A total of 658 Dutch upper-secondary students from three grades (Grade 10, 11 and 12) participated in the national baseline. Data collection took place at 43 schools all over the Netherlands. All the participants of our study were enrolled in a programme forming part of the *VWO* stream (pre-university education). Successful completion of this programme allows the candidates admission to university.

Table 2 presents the distribution of the participants over the three grades and over the schools, and provides information concerning age ($M= 16.95$) and gender (230 males, 428 females) of the participants.

Amongst our participants were 270 grade 10 students (from 34 schools), 271 grade 11 students (from 35 schools), and 117 grade 12 students (from 13 schools). The number of participants in grade 12 is remarkably lower compared to the other two grades. Because of the heavier workload and central exams in the last year of secondary school, the school board proved to be less willing to impose extra activities on these students.

3.3 Data collection procedure

Data collection took place in two rounds. From April to June 2016 data from grade 10 and 11 were collected; from January to

February 2017 data from grade 12 were collected.

Students participated in the study at their own school in groups of ten to twenty students during regular school hours. Data collection was led by two researchers on the project or two trained research assistants. Laptops were provided by the research team. Keystroke logging software Inputlog was installed on the laptops, as were folders with the task sets, including task instructions and sources texts in PDF format, and filling tasks.

Students were first informed of the goal and procedure of the study. After reading and signing the consent forms, students were walked through the synthesis task instructions so they knew what the writing tasks would entail. Instructions included: (1) a short explanation on what a synthesis text is, (2) a short explanation on the characteristics of an argumentative/ informative synthesis text, dependent on the task at hand, (3) instructions on how to deal with the sources, (4) instructions on the audience they had to keep in mind for their text, (5) instructions on style, (6) instructions on text length, and (7) instructions on time. Appendix B presents the instructions in detail. Students had the opportunity to ask questions if the instructions were unclear to them. After that, they also received a short introduction on the use of Inputlog.

Once all students were familiar with the task instructions and the use of Inputlog, they opened the sources on their laptop (without reading) belonging to the version of the first synthesis task assigned to them. The students were instructed to use only the provided sources for their text. Internet use was not allowed. Moreover, participants were instructed to write in the Inputlog document only. Because we wanted to log their complete

Table 2
Distribution of participants over the grades and over the schools

Grade	Schools (N)	Participants (N)	Males/ Females (N)	Age (M)
Grade 10	34	270	84/ 186	15.68
Grade 11	35	271	111/ 160	16.75
Grade 12	13	117	35/ 82	17.35
Total	43*	658	230/ 428	16.59

* Note: In 33 out of 43 schools, students from at least two grades participated.

writing process, they were not allowed to make notes on paper. Students then made sure Inputlog started recording their writing process and had 50 minutes to carry out the task.

After finishing their first text, students stopped the recording of Inputlog. When students finished earlier than the given time, they had to work on one of the so-called filling tasks. These filling tasks were created to keep the students occupied and to make sure that their peers who were still writing would not feel pressurised to rush. After a short break, students carried out the second synthesis task of their task set, again in 50 minutes while recording their writing process with Inputlog. After writing the first two texts, students were given a lunch break. Upon returning in the classroom, they filled in the questionnaire on writing perspectives. Then, the students wrote two more texts, thereby carrying out the third and fourth task of their task set.

3.4 Instruments

Synthesis tasks

Task construction. Given that synthesis writing tasks are rather diverse, the tasks used for this study were diverse too. Creating a variety of synthesis tasks enabled us to draw conclusions about students' general synthesis writing competence instead of for one specific synthesis task. We implemented four different topics, of which the number of sources varied. For all four topics, eight different variants of the task were constructed (see Appendix A) to enable generalisation to a wide range of synthesis tasks. The various versions differed with regard to three relevant task features: (1) the genre of the synthesis text students were asked to write (argumentative synthesis/ informative synthesis), (2) the relation between the source texts (complementary/ contradictory), and (3) the amount of irrelevant information in the source texts (low/ high). When constructing the tasks, we made schematics of the different versions of the sources and how the sources relate to each other. Task construction was done by two researchers on the project, this was then discussed in a team of four. Based

on the discussion, the task construction was then adapted.

Topics. The tasks used for this national survey covered four different topics. These topics were situated in four different interest areas, corresponding to the four study profiles in the upper grades of Dutch pre-university secondary education: Nature & Health (topic: food additives), Nature & Technology (topic: self-driving cars), Culture & Society (topic: the human-wildlife conflict in Africa), and Economy & Society (topic: the pay gap). The synthesis tasks were based on three (food additives), four (self-driving cars and the pay gap), or five (the human-wildlife conflict in Africa) source texts. By varying the number of sources, we addressed the task diversity as the number of sources may have an impact on the process of selecting information. The total number of words across the sources was kept roughly equal for the four topics (and in all task versions), regardless of the total number of sources for that topic. Within each topic the type of sources varied (e.g., newspaper articles, research reports, etcetera), this for all task versions. Amongst the sources of each topic was one source that included numerical information in the form of a table or a graph.

Genre. The informative/ argumentative genre distinction was based on the fact that writing a synthesis requires structuring the information from the sources around a central theme for a communicative purpose, which affects text structure (Bazerman, 1994; Feddema & Hoek, 2018; Swales, 1990). In other words, writing an argumentative synthesis text may require another process of structuring information compared to an informative synthesis text.

Relation between sources. We opted to vary the relation between the source texts in the task construction as this impacts the crucial skill of integrating information. The integration process entails comparing and contrasting the sources; and this activity is influenced by the complementary or conflicting character of the sources.

Amount of relevant source elements. As selecting relevant information from the sources is a required subskill for synthesis

writing, also the amount of irrelevant information in the sources was included as a variable of task variation.

Task set design. We assessed students' writing performance with four tasks as previous studies have shown that more than one task is needed to get a valid and reliable view of a student's writing skills (Schoonen, 2005; Van den Bergh, De Maeyer, Van Weijen, & Tillema, 2012). Each participant wrote four synthesis texts: one on each topic. Task sets were constructed in such a manner that each student wrote two argumentative and two informative synthesis texts in total, of which two texts were based on complementary sources and two were based on contradictory sources, and two synthesis texts were based on sources with little irrelevant information and two were based on sources with a considerable amount of irrelevant information. Task sets were assigned randomly to students. The order of topics was fixed within the school for practical reasons (i.e., otherwise students could inform each other of the different topics during the breaks). Thus, at any given moment, all students from one school were writing about the same topic, but while one student wrote an argumentative synthesis based on complementary sources with little irrelevant information, another student could be writing an informative synthesis based on contradictory sources with a lot of irrelevant information. To avoid an order effect on the quality of the students' texts, the topic order varied randomly over schools (Van Steendam & Bouwer, 2018).

Writing processes

Students wrote their texts on laptops on which keystroke logging software Inputlog (Leijten & Van Waes, 2013) was installed (for more information see the website www.inputlog.net). Inputlog registers mouse movements, keystrokes and window switches. It also offers various types of analyses on the keystroke logging data. Given that Inputlog runs in a familiar word processing environment, it enables us to register the writing process rather unobtrusively.

Students' perspectives on writing questionnaire

The participants filled in a questionnaire in which we enquire after their perspectives on several writing aspects. The questionnaire is based on four validated questionnaires used in previous studies on writing, namely (1) writing apprehension (Rijlaarsdam & Schoonen, 1988), (2) writing beliefs (White & Bruning, 2005), (3) self-efficacy (Braaksma, 2002), and (2) writing process style (Kieft et al., 2007).

First, the writing apprehension questions measure the participants' attitudes towards writing on three levels: cognitive (confidence in one's own writing abilities), affective (writing appreciation) and evaluative (fear of evaluation). Secondly, the questionnaire on writing beliefs contains two scales: transmission (writing seen as a way to transmit knowledge) and transaction (writing seen as a way to transform knowledge by incorporating personal knowledge). Thirdly, the self-efficacy scale enquires after the students' belief in their own synthesis writing abilities. We added a few questions measuring more specific synthesis-related writing abilities to the original questionnaire (for example, *I can select relevant information from different sources when writing a text*). The last part of the perspectives on writing questionnaire contained questions concerning writing process style. These questions measure the participants' levels of planning and revising.

The validity and underlying scales of the various perspectives on writing questionnaires were analysed via factor analyses. Table 3 provides an overview of the scales incorporated into the writing perspectives questionnaire used in this study. It shows the various components incorporated in each scale, the number of items, the item consistency and exemplary items.

In the case of the writing apprehension questionnaire, first a Confirmatory Factor Analysis (CFA) was carried out, given that we used the original questionnaire by Rijlaarsdam and Schoonen (1988). CFA was used to verify if the factors of the original instrument fit our data. The fit indices showed that this model did not fit the data ($\text{cfi} = .804$,

Table 3
Overview of the students' perspectives on writing questionnaire

Scale and components	Number of items	α	Exemplary item
Writing apprehension			
Cognitive	5	.81	When writing a text, I often feel I'm not doing a good job.
Affective	5	.90	I enjoy putting my thoughts on paper.
Evaluative	5	.76	I don't like it when peers read my text.
Writing beliefs			
Transmission	7	.73	The key to successful writing is accurately reporting what authorities think about the subject.
Emotional engagement	5	.74	Writing is a process in which many different emotions play a role.
High amount of revision	3	.60	Writing entails the constant revision of the text to improve what is already written down.
Cognitive engagement	3	.80	Writing helps me to better understand things I'm thinking about.
Self-efficacy			
Dealing with sources	5	.87	I can select relevant information from the sources to write my text.
Language use	3	.77	I can make use of a varied sentence structure and word choice when writing my text.
Concise writing	3	.85	I can write a text without repetition.
Text structure	5	.87	I can structure my text in paragraphs.
Integration of the sources	3	.81	I can relate the information from the different sources in my text.
Elaboration of the sources	2	.70	I can write a source-based text that is clear to someone who did not read the sources.
Writing style			
Preplanning	5	.72	Before I start to write my text, I always make a scheme.
Post-draft revision	5	.74	When I reread and rewrite my text, the content can change a lot.
Short production cycles	4	.72	From time to time I pause writing to revise my text.
Difficult idea generation	4	.72	When writing, I experience difficulties ordering my thoughts.

tli= .788, rmsea= .087, srmr= .092). Consequently, a random portion of the data was explored via Exploratory Factor Analysis (EFA) with oblique rotation. Based on the Kaiser criterion and the scree plot, three factors were identified of which the content relates to the three factors of the original instrument. However, many of the items had rather low factor loadings (< .45). In a next step, we selected the five items with the highest factor loadings for each scale. This model was cross-validated via CFA on the second portion of the data. This resulted into a good fit model, if we take into account two error-covariances. The internal consistency of the three five-item scales is satisfactory (cognitive scale α = .81, affective scale α =

.90, evaluative scale α = .76).
 The fit indices of the CFA on the writing beliefs instrument by White and Bruning (2005) showed that the model did not fit our data (cfi= .730, tli= .692, rmsea= .089, srmr= .077). So, EFA with oblique rotation was carried out on a random portion of the data. Based on the Kaiser criterion, the scree plot and parallel analysis, four factors were identified. The first scale contains seven items related to the transmission idea of writing. The second scale consists of five items related to the idea of writing as a process with emotional engagement. Thirdly, three items related to the idea of writing as a process with a high amount of revision. And the last scale (cognitive engagement) contains

three items related to the idea of writing as a manner to order one's thoughts. For three of the four scales, the internal consistency is good (transmission $\alpha = .73$, emotional engagement $\alpha = .74$, cognitive engagement $\alpha = .80$). Only for the revision scale, Cronbach alpha is low ($\alpha = .60$). Therefore we decided not to take into account this scale in further analyses on the dataset. Our findings are in line with remarks of White and Bruning (2005), who indicated that their transaction scale contained items related to emotions, cognition and revision.

The self-efficacy questionnaire used in our study consists not only of items of the original instrument (Braaksma, 2002) but also of additional items measuring students' self-efficacy in synthesis-specific actions. Therefore an EFA with oblique rotation was carried out on a random part of the data. Depending on the criterion, this resulted in a 1-factor model (based on scree plot), a 2-factor model (based on Kaiser criterion), or a 6-factor model (based on parallel analysis). Contentwise, a 1-factor model is less interesting than a multi-factor model. In a next step, the 2- and 6-factor models were tested via CFA on the second random part of the data. Fit indices and AIC value indicated that the 6-factor model had a better fit ($cfi = .901$, $tli = .881$, $rmsea = .088$, $srmr = .065$, $AIC = 24885.26$) compared to the 2-factor model ($AIC = 23492.8$). Moreover, the internal consistency of each of the six scales is adequate. The scales measure the students' self-efficacy on six aspects: dealing with the sources (reading and selecting information) (five items, $\alpha = .87$), language use (three items, $\alpha = .77$), concise writing (three items, $\alpha = .85$), text structure (five items, $\alpha = .87$), integration of the sources (three items, $\alpha = .81$), elaboration of the sources (two items, $\alpha = .70$).

The last questionnaire, measuring the students' writing style, was based on Kieft et al. (2007). EFA with oblique rotation was used on a random portion of the data. The Kaiser criterion, scree plot and parallel analysis all suggest a 4-factor model. In a next step, the fit of this model was tested on the second portion of the data via CFA. To further improve the model, two items were deleted as they correlated with variables from another

scale. When estimating this model on the complete dataset, the good fit of the model is confirmed ($cfi = .939$, $tli = .928$, $rmsea = .043$, $srmr = .055$). Cronbach's alpha indicates a good internal consistency for each of the four scales: preplanning (five items, $\alpha = .72$), post-draft revision (five items, $\alpha = .74$), short production cycles (four items, $\alpha = .72$), and difficult idea generation (four items, $\alpha = .72$). The preplanning scales measures the degree to which the writer makes a plan before starting to write. The post-draft revision scale indicates the degree to which the writer writes a first complete draft without much revision. The thirds scale measures the degree to which the writers produces in short cycles, revising throughout the process. And the difficult idea generation scale measures the degree to which the writer finds it hard to put things on paper.

3.5 Text quality rating procedure

Assessment method

A total of 2310 synthesis texts was rated by means of a rating scale with benchmark texts. Benchmark rating is a rating procedure in which texts are rated holistically by comparing them to a set of benchmark texts that represent particular points on a text quality scale. Our rating scale contained five benchmark texts at intervals of 1 *SD* (a first benchmark representing a score of - 2 *SD*, a second benchmark with a score of -1 *SD*, an average benchmark, a fourth benchmark scoring +1 *SD*, and a final benchmark of +2 *SD*). All benchmark texts were given an arbitrary score (50 - 75 - 100 - 125 -150).

The benchmark rating procedure was used in previous writing studies (Blok, 1986; Bouwer et al., 2018; De Smedt et al., 2016; Knospe, 2017; Limpo & Alves, 2017; Rietdijk et al., 2017; Rijlaarsdam, 1986; Tillema et al., 2013) as it has several advantages. First, the comparison element facilitates the rating as comparing texts is easier for the rater than assigning a single score (Lesterhuis et al., 2016). Moreover, it increases the validity of holistic rating (Pollitt, 2012) by providing benchmarks accompanied by an explanation of the different criteria included in the global judgement. Thirdly, the raters will be less likely to adapt their judgement during the

writing process as the benchmarks serve as fixed reference points (Bouwer, Koster, & Van den Bergh, 2016). In this way both the effect of sequence and the effect of norm shifting are prevented (Pollmann et al., 2012).

Rating scale construction

We based the rating scale with benchmark texts on the assessment of a random subsample of 150 argumentative and 150 informative synthesis texts on one topic (human-wildlife conflict) with D-PAC, an online tool for comparative judgement (Lesterhuis et al., 2016). The comparative judgement method is based on the assumption that comparing two performances to one another is easier for the rater than assigning a score to one product. The two genres were evaluated in separate assessments as previous research has shown that the textual genre influences performance (Bouwer, Béguin, Sanders, & Van den Bergh, 2015). The (2x 150) synthesis texts were rated on four important synthesis quality aspects separately: (1) relevance and correctness of the information, (2) integration of the sources, (3) coherence and cohesion, and (4) language use), and also got a global judgement. In other words, the same 2x 150 synthesis texts were rated by different groups of raters, each group rating a specific aspect or giving a holistic score. So, the synthesis texts were rated in ten different assessments (five different assessments for each of the two genres). In total, 37 raters were involved. On average, each synthesis text was compared 13.60 times. This led to a rank-order from the lowest to the highest scoring text for each of the ten assessments. The reliability was acceptable to good (SSR reliability coefficient ranging from .60 to .76).

Based on these rankings, we selected benchmark texts to build two rating scales, one for the argumentative synthesis texts, one for the informative synthesis texts. For each rating scale, five benchmark texts were selected (-2 *SD*, -1 *SD*, average text, +1 *SD*, +2 *SD*). In the first instance, we selected texts based on their global score (holistic judgment). Misfit texts, texts on which the scores of the various raters differed

significantly, were not taken into account as they were not considered clear benchmarks. Then we further reduced the selection by selecting those texts with not only a global score approximating the five benchmarks, but also the scores for the four different quality aspects. In a final step, the texts selected were discussed by two researchers and the most representative texts were chosen as benchmarks. See Appendix C for an overview of the various scores of the benchmark texts. Clarifications on each of the four quality aspects for each of the benchmark texts were included as annotations in the final rating scale (for an example see Appendix D).

Rating procedure

The total sample of 2310 synthesis texts was rated with the benchmark scales we constructed. Previous research (Bouwer et al., 2016) showed that the same benchmark scale can be used for rating different writing tasks, at least when texts are written in the same genre. Thus, all four topics were rated by means of these two genre-specific scales (i.e., for the informative and argumentative genre). Raters were instructed to compare the students' texts to the benchmark texts. Any score could be given (thus, also scores below and above the benchmark scores were accepted). We asked the raters to include four criteria in their global judgement: (1) relevance and correctness of the information, (2) integration of the sources into a new text with its own structure and overarching theme, (3) coherence and cohesion, and (4) language use. We based these criteria on previous research on synthesis writing (Boscolo et al., 2007; Mateos et al., 2008; Mateos & Solé, 2009; Solé et al., 2013).

Raters

A design of overlapping rater teams was applied (Van den Bergh & Eiting, 1989). This procedure entails that the texts to be rated were split randomly into several subsamples and each rater rates three subsamples according to a prefixed overlapping design. In this way, every text was rated by a jury of three raters.

The argumentative texts were rated by 24 raters using the rating scale with argumentative benchmark texts. Another 24 raters assessed the informative synthesis texts using the rating scale with informative benchmark texts. Every individual rater rated only one genre of synthesis texts and only one topic; this was done in order not to complicate the job of the rater as he/she had to take into account the task-specific sources when assessing the texts.

Part of the raters were Dutch teachers, part of the raters were master's students and PhD researchers enrolled in a language-oriented study. Prior to the actual assessment of the texts, all raters were given a training in small groups. They received the rating scale and a set of 5 texts in order to practice. The assessment method and the rating of the exemplary texts were then discussed in groups of two to three people via Skype sessions with two researchers on the project.

After the training, the raters received a set with 150 texts. They were given three to four weeks to complete the assessment. In total, it took them approximately eight hours to complete the assessment. Raters received a financial reward for their cooperation.

The average jury rater reliability was .65 ($\rho = .65$, $se = .08$). The final score per text consisted of the mean of the three scores given by the raters.

3.6 Process data preparation

Filtering and recoding of Inputlogfiles

Prior to running the analyses, the Inputlog data were prepared by using the time filter and source recoding functions of Inputlog. First, the time filter removed possible clutter at the end of the writing process (e.g. actions to stop the Inputlog recording). All the writing process files were filtered at the last key, that is, we considered the moment at which the last character was typed as the end of the writing process. Secondly, the source recoding function was used to group several sources identified by Inputlog into one of the following source categories: a given source text, the synthesis text written by the student, and off-task sources (e.g. internet sources).

Process measures

All writing processes were analysed using Inputlog version 8.0.0.5. Based on the data generated by the Inputlog analyses, we created 11 process indicators, which give information on five main synthesis writing process aspects, namely general time usage, production, pausing, revision and source use. The selection of process variables was guided by two principles, namely (1) interpretability (the variables are interpretable in the context of one of the five main writing process aspects), and (2) clarity (the indicators have to be clear and straightforward, which will

Table 4
Overview of the selected writing process variables

Process aspect	Process variable	Overall process	Three intervals
Time usage	Total process time	✓	
	Proportion of time in sources		✓
	Proportion of active writing time (during production)		✓
	Proportion of pause time (during production)		✓
Production	Number of keystrokes typed	✓	
	Number of keystrokes per minute		✓
Pausing	Number of pauses per minute (during production)		✓
	Mean pause time (during production)		✓
Revision	Produced ratio (= number of characters in the final text divided by the total number of characters produced during the process)	✓	
Source use	Number of transitions per minute between the sources		✓
	Number of transitions per minute between the synthesis text and the sources		✓

allow transfer to educational contexts, such as feedback on the writing process).

Each writing process was divided into three equal intervals: beginning, middle and end (Breetvelt et al., 1996). We took into account the timing in the process (i.e. interval) for eight process variables. For the other process variables it was not possible to calculate the interval variables based on the Inputlog data. Thus, with three process variables giving information on the overall writing process and eight process variables providing interval-related information, a total of 27 process variables were available per text. Table 4 provides an overview of the process variables used in this study. Most of these process variables are relative measures (e.g. proportions and actions per minute). These relative measures allow us to compare the writing processes (as some students finished earlier than the given 50 minutes time on task) and to generalise the findings.

3.7 Analysis

Text quality

The structure of our data is rather complex: text quality scores are nested within students; and students are nested within schools. As students wrote several tasks, the text quality scores are also dependent of the task. Moreover, the design of our study implies that students and tasks were crossed. Given this hierarchical and cross-classified structure, data were analysed using mixed-effect modelling. This allowed to capture the complex data structure and to estimate the variances between schools, between students, between tasks and an error variance component. The use of mixed-effect modelling reduces the probability of Type-I errors; moreover, because both student and task characteristics can be included as independent variables, mixed models usually allow for more rich interpretations (Hox, 2002; Quené & Van den Bergh, 2008).

Four models were built to examine the effect of grade, text genre, and gender on the students' writing performance. Starting with a null model that did not contain any explanatory variables, only random effects, we successively added explanatory variables:

- Null model: without any explanatory variables, only random effects (participant, school, task)
- Model 1: main effects of gender, grade and genre
- Model 2: main effects + interaction effect of grade and gender
- Model 3: main effects + interaction effect of grade and genre

To test the difference between the several models, we applied the Likelihood Ratio Test. Chi-square goodness of fit test was then used to determine the model with the best fit (Curran et al., 2010).

Writing process

The writing process data were also analysed using mixed-effect modelling as the various writing process variables are nested within students and within tasks. Thus, both student (grade and gender) and task characteristics (genre) were taken into account when analysing the writing process variables. We tested the same four models as for text quality.

To facilitate the interpretation of the results, we opted to work with standardised values for all writing process variables. Z-scores for all process variables were calculated.

Prior to conducting the analyses, several checks were performed to assure the accuracy of the Inputlog data. It is important to note that keystroke logging data should be handled with care because of possible technical failures or actions of the students that can distort the view on the writing process. A first check was performed on a variable not included in our final analyses: *proportion of time in other sources*. This is the time students spent in sources that were not the sources we provided them with, nor the word document they were writing their synthesis text in. Actions like checking the clock, going to the computer's main menu etcetera were coded as "other sources"; in these cases, off-task time was limited. However in some cases we noted that the value for *proportion of time in other sources* was rather high. This was the case when students for example were

performing the wrong task, or were consulting the internet. After revising some cases, it was decided to set the threshold on 0.10. So, cases ($N = 67$) in which more than 10% of the process was spent in “other sources”, and thus off-task, were not included in any of the analyses.

Secondly, in the case of the variable *mean pause time*, we noticed that several cases ($N = 40$) had a missing value in the first interval of the process. Due to a technical error, this variable was not processed by the Inputlog analysis. These cases were excluded from the process analyses.

After performing these two checks to assure the validity of the Inputlog data, the distribution of the data was controlled. Visual inspection via histograms showed that the variables *number of transitions per minute between the sources* and *number of transitions per minute between synthesis text and sources* were not normally distributed. Log-transformation was applied to these two variables so as to approach normal distribution. Analyses were carried out with the log-transformed variables.

Students’ perspectives on writing
Mixed-effect modelling was used to analyse the development in students’ perspectives on writing over the grades as students are nested

within grades and within schools. Also gender was taken into account as a student characteristic. We tested a null model (with school as a random effect), a model with the main effects of grade and gender (model 1) and a model in which we added the interaction between gender and grade (model 2).

4 Results

4.1 Text quality

We compared four models to determine whether the quality of the students’ synthesis texts is dependent on gender, grade or synthesis genre; moreover, various interaction effects were examined. The model fits and comparisons are shown in Appendix E. Appendix F shows the parameter estimates for the best fitting model (model 3, $\chi^2(2) = 13.01$, $p = .001$). The interaction effect between grade and genre is plotted in a graph (Figure 1). Post-hoc tests showed that the average writing score differed significantly between the three grades for both argumentative and informative synthesis texts. In grade 10, students scored on average 10.37 points (equivalent to .54 *SD*) lower than in grade 11 ($p < .001$), and in grade 12 they scored 4.35 points (equivalent to .23 *SD*) higher than in grade 11 ($p = .052$) in the case

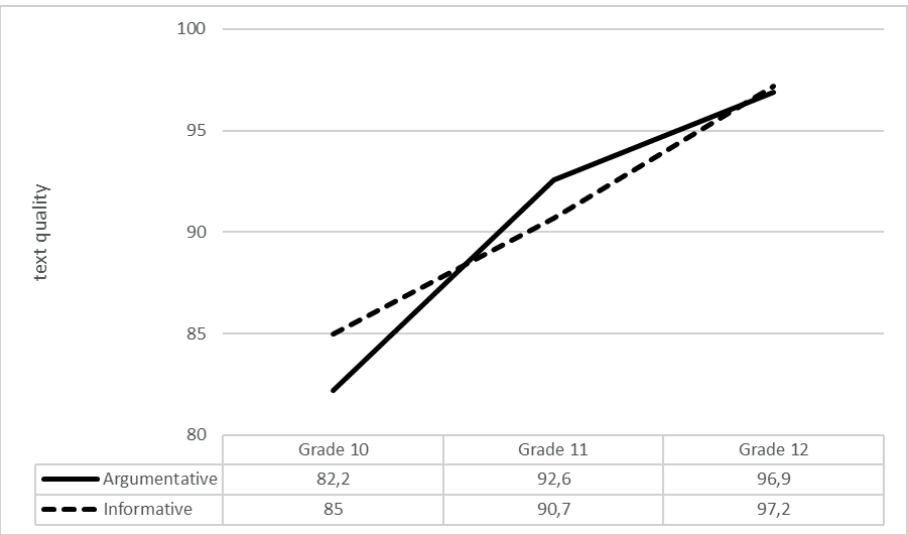


Figure 1. Interaction effect of grade and genre for text quality.

of the argumentative texts. For the informative genre, the grade 10 students scored 5.67 points (equivalent to .32 *SD*) lower than the grade 11 students ($p < .001$), and the grade 12 students scored 6.51 points (equivalent to .37 *SD*) higher than the grade 11 students ($p = .002$). The interaction between grade and genre implies that the growth between grade 10 and grade 11 differs according to genre, with the argumentative genre increasing more than the informative genre.

Random effects showed that only a small proportion of the variance in text quality could be attributed to the school ($ICC = .04$). So, the school had little effect on the students' writing performance.

4.2 Writing process

To map the development of the writing process over the grades and to test the effect of genre and gender, we analysed several writing process variables obtained via keystroke logging, grouped into five process aspects: (1) general time usage, (2) production, (3) pausing, (4) revision, and (5) source use. In Appendix E the model fits and comparisons for all variables can be found. Based on these model comparisons, the best-fitting model was identified. The parameter estimates for the best fitting model of each variable can be found in Appendix F. The variance in writing process attributable to schools varied between an ICC value of .00 and .06. So, the school to which the students belonged, had no to little effect on their writing process.

Table 5 presents an overview of the effect for each of the 27 writing process variables. This table shows:

Whether there was an effect of grade, gender, genre, or an interaction effect; in other words, it shows which model is the best fitting model.

The specific contrasts and their direction; in other words, it shows how the three grades, two genders and two genres are positioned against each other.

The size of the effect (expressed in standard deviation) and the significance (p -value); in other words, it shows how big the contrast is between the three grades, two

genders and two genres.

In addition to the table, the effect of grade, gender and genre are briefly described for the five process aspects and their underlying process variables in the following sections (sections 4.2.1 to 4.2.5).

General time usage

First, we looked at the general distribution of the main actions of the writing process by analysing the total duration of the writing process, the proportion of time spent at actively writing the synthesis text, the proportion of time spent at pausing during production and the proportion of time spent at reading the sources.

Total process time. Model 3 with an interaction effect between grade and genre proved to be the best fitting model for the total time on task ($\chi^2(2) = 11.68, p = .003$). When writing an argumentative synthesis text, grade 10 students spent less time on task than students in grades 11 and 12. For the informative genre, the grade 11 students' total process time was shorter than that of the grade 12 students.

Proportion of time in sources. The proportion of time the students spent in the sources was observed for each of the three intervals. For the first interval, the model with the main effects only (model 1) resulted the model with the best fit ($\chi^2(4) = 25.54, p < .001$). The students in the last year of upper-secondary education spent a significantly lower amount of time in the sources during the beginning of the writing process. The gender effect implies that boys spent a higher proportion of time in the sources during the first interval. And thirdly, concerning genre, students spent more time in the sources during the first phase when doing an informative compared to an argumentative task.

For the proportion of time in sources during the second interval, model 2 was the model with the best fit ($\chi^2(2) = 11.01, p = .004$). This implies an interaction effect between grade and gender. First, the interaction effect means that the differences in proportion of source time between the three grades were only significant for girls.

Table 5
Overview of the effects of grade, gender and genre for the writing process variables: best-fitting model, contrasts and effects

Process variable	Model 0	Model 1			Model 2	Model 3	Contrasts	Effect (estimates in SD) and significance
	Null model	Grade	Gender	Genre	Grade x Gender	Grade x Genre		
Total process time						x	ARG: 10 < 11, 12 INF: 11 < 12	10 - 11: -.21*, 10 - 12: -.40* -.26*
Proportion of time in sources interval 1		x	x	x			10, 11 > 12 F < M ARG < INF	11 - 12: +.25** -.14* -.24*
Proportion of time in sources interval 2					x		F: 10 > 11 > 12 10: F > M 12: F < M	10 - 11: +.20*, 11 - 12: +.30* 0.24* -.32*
Proportion of time in sources interval 3		x		x			10 > 11 > 12 ARG < INF	10 - 11: +.016*, 11 - 12: +.30* -.30**
Proportion of active writing time (during production) interval 1		x	x	x			10, 11 < 12 F > M ARG > INF	11 - 12: -.30** +.22** +.18*
Proportion of active writing time (during production) interval 2		x					11 < 12	-.26**
Proportion of active writing time (during production) interval 3	x							
Proportion of pause time (during production) interval 1			x				F < M	-.15*
Proportion of pause time (during production) interval 2		x		x			10 < 11 ARG > INF	-.14* +.21**
Proportion of pause time (during production) interval 3		x		x			10 < 11 ARG > INF	-.16* +.14**
Number of keystrokes typed		x	x				10 < 11 < 12 F > M	10 - 11: -.31**, 11 - 12: -.46** +.31**
Number of keystrokes per minute interval 1		x					10 < 11 < 12	10 - 11: -.21**, 11 - 12: -.24*
Number of keystrokes per minute interval 2		x		x			10 < 11, 12 ARG > INF	10 - 11: -.19* +.14*
Number of keystrokes per minute interval 3						x	10: ARG > INF 12: ARG > INF	+.25** +.19*
Number of pauses per minute (during production) interval 1	x							
Number of pauses per minute (during production) interval 2					x		F: 10 < 11, 12 10: F < M	10 - 11: -.22*, 10 - 12: -.30* -.22*
Number of pauses per minute (during production) interval 3						x	ARG: 10 < 11 INF: 10 < 11, 12 10, 11: ARG > INF	-.16** 10 - 11: -.25**, 10 - 12: -.44** +.21**
Mean pause time (during production) interval 1		x		x			10 < 11 ARG > INF	-.12* +.09*
Mean pause time (during production) interval 2				x			ARG > INF	+.27**
Mean pause time (during production) interval 3	x							
Produced ratio						x	ARG: 10 > 11 > 12 INF: 10, 11 > 12 11, 12: ARG < INF	10 - 11: +.26**, 11 - 12: +.48** 10 - 12: +.55**, 11 - 12: +.45** -.11*, -.14*
Number of transitions per minute between the sources interval 1				x			ARG < INF	-.10*
Number of transitions per minute between the sources interval 2					x		12: F < M	-.28*
Number of transitions per minute between the sources interval 3					x		M: 10 < 11 10: F > M 11: F < M	-.16** +.11* -.10*
Number of transitions per minute between synthesis text and sources interval 1		x	x	x			10 < 11 < 12 F > M ARG < INF	10 - 11: -.10*, 11 - 12: -.17* +.27** -.26**
Number of transitions per minute between synthesis text and sources interval 2				x			ARG < INF	-.28**
Number of transitions per minute between synthesis text and sources interval 3			x	x			F < M ARG < INF	-.08* -.29**

Note: * significant at the p < .050 level, ** significant at the p < .010 level

The higher the grade, the less time female students spent in the sources. Secondly, the difference between boys and girls was significant in two grades: in grade 10, girls spent more time in the sources compared to boys, while in grade 12 they spent significantly less time.

For the last interval, the model with only the main effects (model 1) resulted the best ($\chi^2(4) = 50.91, p < .001$). Regarding grade effect, results indicated that the higher the grade, the lower the proportion of time spent in sources. The second main effect is the genre effect: the proportion of time in the sources during the third interval was significantly higher for informative texts than for argumentative texts.

Proportion of active writing time. The proportion of active writing time indicates the amount of time the writer spent in each interval at the actual production of the text. For the first interval, model 1 had the best fit ($\chi^2(9) = 29.64, p < .001$). There were three main effects: grade, gender and genre. First, the effect of grade: students from grade 12 had a significantly higher proportion of active writing time in the first phase of the process compared to students from grades 10 and 11. Secondly, gender had an effect: the proportion of active writing time was lower in the case of boys. Thirdly, genre had an effect: for the informative tasks, the active writing time was lower than for the argumentative task.

Also in the second interval, the model with only the main effects (model 1) proved to be the best fitting model ($\chi^2(9) = 9.40, p = .052$). The effect of grade was significant: in the middle phase of the process, grade 12 students spent a higher amount of time at actively writing their text than grade 11 students.

For the proportion of active writing time in the third interval, model 1 was not significantly better than the null model. So, nor grade, nor gender, nor genre had an effect on the active writing time of the last phase of the writing process.

Proportion of pause time during production. Pauses during production are periods of two seconds or more, spent in the word document, when no activity is registered. The proportion of pause time was

analysed for each of the three intervals. For the first interval, model 1 resulted the best fitting model ($\chi^2(9) = 9.51, p = .049$). Grade and genre effect were not significant. Only gender proved to have a significant effect. In the beginning of the writing process, the proportion of pause time was significantly higher for boys compared to girls.

Also in the second part of the process, it was the first model that had the best fit ($\chi^2(9) = 24.78, p < .001$). Both grade and genre had a significant effect on the proportion of pause time in the middle of the process. Students in grade 10 paused significantly less than grade 11 students. Moreover, the proportion of pause time was lower for the informative genre than for the argumentative genre.

Model 1 was also the best fitting model for the proportion of pause time in the third interval ($\chi^2(9) = 17.54, p = .002$) with an effect of both grade and genre. Similarly as in the previous writing process phase, the proportion of pause time was lower in grade 10 and in the case of informative tasks.

Production

For the second key writing process aspect, production, we took into account two process measures. First we analysed the total amount of keystrokes typed during the whole process; in other words, all the characters that the writer produced while working on the synthesis text. Secondly, we also took into account the (fluency of) production in each of the three writing process phases as this may indicate processing difficulties during writing (Olive & Kellogg, 2002). Production fluency was measured by the number of keystrokes per minute in each of the three process intervals.

Number of keystrokes typed. Model 1 ($\chi^2(9) = 75.77, p < .001$) resulted the best fitting model for the total number of keystrokes typed during the process. There was both a grade and a gender effect. There was an increase of total keystrokes typed over the grades. And the text production was on average less fluent in the case of boys compared to girls.

Number of keystrokes per minute. Model 1 proved to be the best fitting model ($\chi^2(9) =$

21.78, $p < .001$) for the number of keystrokes in the first part of the writing process. There was a main effect of grade. The higher the grade, the more fluent students wrote in the first interval.

For the number of keystrokes per minute in the second interval, model 1 was the model with the best fit ($\chi^2(9) = 18.99$, $p < .001$). There was a grade effect: grade 10 wrote less fluently than grades 11 and 12 in the middle of the writing process. Moreover, there was also a significant effect of genre, given that students produced less keystrokes per minute when writing an informative synthesis text, than when writing an argumentative text.

Model 3 proved to be the best fitting model for the third interval ($\chi^2(11) = 6.49$, $p = .039$). In the last episode of the process, an interaction between grade and genre was observed. Both within grade 10 and grade 12, the number of keystrokes per minute was higher in the case of the argumentative genre compared to the informative genre.

Pausing

The third key writing process aspect under study was the pausing behaviour during production. Besides time spent at the sources and at actively writing the text, there is also an amount of time spent at pausing. This pausing time can be related to thinking time: students plan what to write next, they trying to generate ideas, they reread what is already written, or they are simply stuck. To map the pausing behaviour, we studied two variables: the number of pauses per minute and the mean pause time. These give us information on how many times writers paused during production (pausing frequency) and the length of the pauses (pause duration). The temporal distribution was taken into account as these pause-related variables were analysed for each of the three process intervals.

Number of pauses per minute (during production). In the first interval, there was no effect of grade, genre nor gender on the number of pauses per minute.

During the second interval results indicated an interaction effect of grade and gender ($\chi^2(11) = 6.76$, $p = .034$, Model 2). For boys, there was no effect of grade. For

girls though, there was a difference between grade 10 on the one hand and grades 11 and 12 on the other hand. More specifically, the younger female students paused less frequently. There was no significant difference between boys and girls, except in grade 10 where girls paused less frequently than boys.

For the last interval of the writing process, model 3 with an interaction effect between grade and genre proved to be the best-fitting model ($\chi^2(11) = 8.23$, $p = .016$). First, regarding differences between the grades, results indicated that there was a difference between grades 10 and 11 for both argumentative and informative genre. More specifically, for both genres the number of pauses during production per minute was lower in grade 10 than in grade 11. For informative tasks there was also a difference between grades 10 and 12, being the number of pauses during production lower in grade 10. Secondly, there were differences between the genres in grades 10 and 11. In these two grades, students paused more frequently when writing an argumentative synthesis text than when writing an informative synthesis text.

Mean pause time (during production). The average time students spent pausing during production in the beginning of the writing process, was significantly affected by grade and genre ($\chi^2(9) = 3.46$, $p = .009$; Model 1). The mean pause time was lower in grade 10 compared to grade 11. Moreover, the mean pause time was lower in the case of an informative synthesis task. Gender had no significant effect on the average duration of pausing time in the first interval.

Also in the middle phase of the process, model 1 resulted the model with the best fit ($\chi^2(9) = 13.65$, $p = .009$). The genre effect implies that the mean pause time was lower in the case of informative synthesis texts. For grade and gender no effect was found.

Model comparison showed that for the last interval, the null model was the best-fitting model (Appendix E). So, there was no effect of grade, gender, nor genre.

Revision

For the fourth writing process aspect, revision, we took into account the *produced ratio*

variable. This variable gives an indication of the overall revision ratio as it consists of the number of characters in the final text divided by the number of characters produced during the process. So, if the ratio has a value of 1, no revision took place as all characters produced during the process ended up in the final text. The lower the ratio, the more revision.

Produced ratio. Revision was significantly affected by an interaction between grade and genre ($\chi^2(11) = 9.60, p = .008$, Model 3). For the argumentative genre, there was a significant difference between all three grades: the higher the grade, the lower the produced ratio. For the informative genre, there was a difference between grade 12 on the one hand and grades 10 and 11 on the other, being the produced ratio lower in grade 12. When analysing the difference between the two synthesis text genres in each grade, results showed there the produced ratio was lower for argumentative tasks than for informative tasks in both grade 11 and grade 12.

Source use

The last process aspect under study was source use. Two variables were selected that map the switching behaviour between the various sources, and between the sources on the one hand and the synthesis text in production on the other hand.

Number of transitions per minute between the sources. For the first interval, the model with only the main effects proved to be the best fitting model ($\chi^2(9) = 9.47, p = .050$, Model 1). Genre had a significant effect. Students switched more between the sources at the beginning of the process when writing an informative text than when writing an argumentative text.

In the case of the transitions between sources in interval 2, model 2 with an interaction effect between grade and gender ($\chi^2(11) = 7.44, p = .024$) had the best fit. In grade 12, girls switched less frequently between the sources than boys.

The interaction effect between grade and gender was also observed in the last interval of the process ($\chi^2(11) = 11.41, p = .003$, Model 2). First, there was a significant

difference between grade 10 and grade 11, but only for boys. In grade 11, boys switched more than in grade 10. Secondly, a difference between boys and girls was observed in two grades. In grade 10, girls switched significantly more between the sources at the end of the writing process. In grade 11, it were the boys that switched more.

Number of transitions per minute between the synthesis text and the sources. In interval 1, model 1 was the model with the best fit ($\chi^2(9) = 56.27, p < .001$). The three main effects were significant. First there was an effect of grade: the higher the grade, the more transitions between synthesis and sources at the beginning of the writing process. Secondly, also gender had an effect: boys switched less frequently than girls. And thirdly, regarding genre effect, we found that there were more switches per minute between the synthesis text and the sources in the case of the informative synthesis genre compared to the argumentative genre.

For interval 2, model comparisons indicated model 1 as the best-fitting model ($\chi^2(9) = 17.37, p = .002$). There was a significant effect of genre as there were more switches between the synthesis text and the sources during informative tasks than argumentative tasks.

For the number of transitions per minute between the synthesis text and the sources in the last interval, model 1 had the best ($\chi^2(9) = 51.57, p < .001$). There was a main effect of gender as boys switched more between the synthesis text and the sources than girls and a main effect of genre as the frequency of switches was higher in the case of informative tasks.

4.3 Students' perspectives on writing

Three models were tested to map the development over the grades and to explore the effect of gender on each of the writing perspectives components of the four scales: writing apprehension, writing beliefs, self-efficacy and writing style.

The model fits and comparisons for all variables can be found in Appendix E; the parameter estimates for the best fitting models in Appendix F.

Inspection of the random effects showed that the ICC of schools varied between .00 and .04 for the majority of the writing perspectives components. However, for the affective component and the preplanning component, the ICC value was higher. So, the variance in how much students like to write is for 9 % attributable to the school the participants belong to; and the variance in preference for preplanning style is for 7% due to school differences.

Writing apprehension

Cognitive. Model 1 ($\chi^2(6) = 8.10, p = .044$) had the best fit. The main effect of gender implied that girls scored .19 SD higher than boys ($p = .023$) for the cognitive component of writing.

Affective. The model with the best fit for the affective writing component was model 1 ($\chi^2(6) = 14.46, p = .002$). Parameter estimates indicated both a grade and a gender effect. In grade 12, students scored .27 SD higher ($p = .031$) than in grade 11, and girls scored .24 SD higher ($p = .003$) than boys.

Evaluative. For the evaluative component, there were no effects of grade or gender.

Writing beliefs

Transmission. For the transmission scale, model 1 had the best fit ($\chi^2(6) = 8.07, p = .045$). The degree to which students see writing as a way to transmit knowledge is lower in grade 12 than in grade 11 (difference of 29 SD, $p = .014$).

Emotional engagement. Model 1 ($\chi^2(6) = 9.44, p = .024$) proved to have the best fit. Girls were .22 SD more emotionally engaged in writing than boys ($p = .008$).

High amount of revision. Model comparisons showed that model 1 was the best-fitting model ($\chi^2(6) = 10.75, p = .013$). There was a main effect of gender as girls scored .16 SD higher ($p = .052$).

Cognitive engagement. Model 1 had the best fit ($\chi^2(6) = 12.72, p = .005$). Parameter estimates indicated an effect of grade and of gender. Grade 12 scored .26 SD higher than grade 11 ($p = .028$), and girls scored .21 SD higher than boys ($p = .012$) regarding writing seen as a way of ordering one's thoughts.

Self-efficacy

Dealing with sources. Model comparisons showed that model 1 ($\chi^2(6) = 8.65, p = .034$) had the best fit. Parameter estimates showed that there was a main effect of gender ($p = .012$), that is, girls scored .21 SD lower on self-efficacy regarding dealing with sources than boys.

Language use. Also for language use self-efficacy, the first model ($\chi^2(6) = 11.74, p = .008$) was the best-fitting one. There was a grade effect: students in grade 12 felt more confident than 11th grade students (difference of .39 SD, $p = .001$).

Concise writing. Though model comparisons showed that the null model had the best fit, parameter estimates indicated an effect of gender ($p = .046$). Girls scored .17 SD lower than boys regarding self-efficacy in concise writing.

Text structure. Model 2 ($\chi^2(8) = 6.20, p = .045$) resulted the best-fitting model. There was an interaction effect between grade and gender. Parameter estimates showed that students in grade 10 scored .32 SD lower on text structure self-efficacy than grade 11 students ($p = .026$); moreover, in grade 11 there was a difference between boys and girls as the girls scored .43 SD lower ($p = .001$). Given that the difference between boys and girls in grade 10 is smaller (.45 SD, $p = .013$), there was no significant difference between the two genders in grade 10.

Integration of the sources. The null model had the best fit according to the model comparisons. However, when looking at the parameter estimates a gender effect was observed. Girls felt .19 SD less confident on the integration aspect ($p = .025$).

Elaboration of the sources. Though the null model had the best fit, parameter estimates indicated a gender effect. Girls felt .22 SD less confident ($p = .008$) than boys regarding elaboration of the sources.

Writing style

Preplanning. Model 1 ($\chi^2(6) = 14.38, p = .002$) was the best-fitting model. First, there was a grade effect: in grade 11, students scored .17 SD higher on preplanning than in grade 10 ($p = .043$). Secondly, there was a

gender effect: girls scored .26 *SD* higher than boys ($p = .001$).

Post-draft revision. Model comparisons indicated model 1 ($\chi^2(6) = 13.67, p = .003$) as the model with the best fit. Parameter estimates showed that there was an effect of gender. Girls' writing style more resembled a drafting to explore or post-draft revision writing style than boys' writing style (difference of .18 *SD*, $p = .027$).

Short production cycles. For the writing style involving writing in short cycles, model 1 had the best fit ($\chi^2(6) = 11.55, p = .009$). There was a main effect of gender. Girls scored .21 *SD* higher for this writing style ($p = .009$).

Difficult idea generation. There were no effects of grade nor gender (null model).

5 Discussion and conclusion

Our general goal was to map synthesis writing in Dutch upper-secondary education via a national survey study. As a national survey study, this study is purely descriptive. It provides an overview of the actual situation: how well do students perform on synthesis tasks?, how do they write synthesis tasks?, and what are their perspectives on synthesis writing? This study is the first one to analyse the writing of a large national sample of pre-university students (*VWO*-stream) in the Netherlands on three levels: the quality of the text, the writing process and students' perspectives on writing. In our analyses, we took into account the effect of three factors on synthesis writing: grade, gender, and task genre. We aimed to describe the development of text quality, writing process and perspectives over the three highest grades of secondary education, and how this differed for argumentative and informative synthesis texts, and for boys and girls. By looking at the effect of three factors (grade, gender, genre) on three aspects of synthesis writing (product, process, perspectives), we offer a fairly complete view on the current state of synthesis writing and a baseline with rich data. This baseline can help shape future intervention studies and classroom practice.

For example, based on the results of this study, motivated decisions can be made for instructional design. The insights from this study can help to differentiate writing instruction regarding grade, gender or task genre.

In this discussion, we first give an overview of the results (descriptive findings). We will also include concrete examples of possible implications for future educational research or practice. In addition, we focus on some methodological aspects that are crucial for national survey studies. Last, we point out some limitations of the study, and possibilities to overcome those in the future.

Regarding the first aim - mapping the effect of grade on synthesis writing - we have found that grade not only has an effect on students' writing performance, but also on their writing process and their perspectives on writing. These results confirmed findings from previous studies regarding grade effects (Graham, 2018; Martínez et al., 2015; Mateos & Solé, 2009). First, the higher the grade, the better the students perform in terms of text quality: the students in the higher grades write better synthesis texts. Secondly, also the way in which students write their text varies over the grades. We have observed changes for various writing process aspects, both for the overall writing process and during several intervals of the writing process. When taking into the account the overall writing process, we see that students in grade 10 (the youngest students in our sample) spend less time on task. Moreover, the younger the students, the less text they produce. Students' revision behaviour also changes over the grades: the higher grade students revise more (for the argumentative genre, the amount of revision increases over the years; for the informative genre, grade 12 students revise significantly more than the two lower grades). We also have observed a development in the three writing process intervals over the grades. In the beginning of the writing process (first interval), students in grade 12 (the highest grade of our sample) spend less time in the sources and more time actively writing their text compared to students in the two lower grades. The higher

the grade, the more the students switched between the synthesis text and the sources. Moreover, the higher the grade, the more fluent the students write in the first interval. Grade 10 students pause for shorter periods of time at the beginning of the process than the students in grade 11. Also during the middle part of the writing process, the effect of grade was observed. The higher the grade, the less time students spend in the sources, though this is only the case for the female students. Compared to grade 11 students, the students in grade 12 spend more time actively writing their text during the second interval. Grade 10 students spend less time pausing during production than students in grade 11. Students in grade 10 produce text in a less fluent way during the middle part of the process, and they pause less frequently (though the latter only counts for the female students) than the two higher grades. In the third and last part of the writing process, the higher grade students spend less time in the sources. Grade 10 students spend less time pausing than grade 11 students. They also pause less frequently than grade 11; and, in the case of the informative task genre, also less than grade 12. Moreover, during the last part of the process, grade 10 students switch less between the sources than grade 11 students, though this only counts for boys. Besides text quality and writing process, also the third aspect under study, the students' perspectives on writing, develops over the grades. We have found that grade 12 students feel more positive towards writing (affective component), they also consider writing less as a mere way of transmitting knowledge, and they are more cognitively engaged in writing compared to the younger students. For self-efficacy, the students in grade 12 differ from those in grades 10 and 11: they feel more confident when it comes to language use and structuring the text.

As a second aim for the national survey, we wanted to map the effect of gender on text quality, writing process and perspectives (Cordeiro et al., 2018). We have not only found an effect of gender on the quality of the synthesis texts, but we have also identified several differences between girls and boys in how they write their text. For the overall

writing process, girls produced more text than boys. When looking at gender differences in the first interval of the writing process, we see that girls spend less time in the sources, more time actively writing the text, and less time pausing during production. Girls also switch more between the sources and their own text at the beginning of the process. In the middle part of the process, girls in grade 10 spend more time in the sources than boys; while in grade 12 they spend less time in the sources than boys. Girls also pause less frequently (though this only counts for grade 10). Girls switch less frequently than boys in the second interval (only in grade 12). Also in the last part of the writing process, there are differences between the two genders. In grade 10, girls switch more frequently between the sources than boys, and in grade 11 they switch less frequently between the sources. Regarding switches between synthesis text and sources, girls switch less than boys in the last part of the process. For the third aspect under study, the students' perspectives on writing, we have found several effects of gender. Girls find writing more cognitively demanding, they feel more positive towards writing, and they are more emotionally engaged in writing than boys. Regarding self-efficacy, girls feel less confident than boys on quite a few aspects: dealing with sources, concise writing, text structure (though the gender difference is only present in grade 11), integration of sources, and elaboration of sources. For writing style it seems that girls have a more outspoken writing style preference as they score higher on three writing styles: preplanning, post-draft revision, and short production cycles.

The third aim of our study was to map the effect of genre (Bouwer et al., 2015) on two aspects of students' synthesis writing. First, there is an effect of genre on the quality of the synthesis texts. Though the text quality improves over the grades for both argumentative and informative genre, we have found that students make more progress between grade 10 and grade 11 for the argumentative genre. Secondly, also for the writing process variables we have found effects of genre. For the overall writing process, we have found that students revise

more when writing an informative text (this only counts for grades 11 and 12, not for grade 10). In the first interval of the writing process, students spend more time in the sources and less time actively writing when working on an informative task. The mean time of the pauses during the first part of the process is longer when students write an argumentative text. Moreover, students switch more between the sources in the first interval when working on an informative task, but they switch more between the synthesis and the sources when writing an argumentative text. During the second part of the writing process, the proportion of pause time is higher for the argumentative tasks, so students pause more while working on an argumentative text. In the last interval of the process, students spend less time in the sources and more time pausing during production when writing an argumentative text. In grades 10 and 12, text production is more fluent in the last part of the writing process of an argumentative text. Also the pausing frequency is higher for argumentative texts (this counts for grades 10 and 11). When working on an informative task, students switch more between the synthesis text and sources at the end of the process, compared to when writing an argumentative text.

This national baseline showed that students' synthesis writing performance increases with regular schooling. Though this is a positive result, it should be noted that, while secondary students are expected to write a synthesis text, instruction on this type of writing is scarce (Van Ockenburg et al., 2018). Moreover, we know that even in higher education, students struggle to write a successful synthesis (Mateos & Solé, 2009). So, developing instruction to help students improve their synthesis writing is important. Classroom practice can build on the insights from the national survey study. This study showed that writing processes develop over the grades; and though our study did not relate the process to the text quality, it is clear that more experienced writers approach the writing process differently. Our findings thus suggest that it is important to take into account writing processes in writing instruction. By

focusing on the process aspects, students will become aware of their writing behaviour. In addition, our findings suggest that information on aspects such as students' self-efficacy and writing styles is valuable to understand students' writing and to guide them in becoming better writers.

The results related to the effects of grade, gender and genre on students' synthesis writing, have some implications for future educational research or practice. First, the results regarding grade effect, could help decide which process aspects to focus on in writing instruction. For example, results seem to indicate that grade 10 students struggle with dealing with the sources (as they spend less time actively writing in the beginning and more time reading the sources at the end compared to the higher grades). An implication for instructional design could be to focus on offering grade 10 students ways to deal with the sources, such as helping them select the relevant information and taking notes (and thus actively write) in the beginning of the process, so that later on in the process they can focus more on the actual writing of the text. Secondly, also the results related to the gender effect could serve as input for instructional design. We know from the present survey that boys and girls show differences in their writing process, and in their perspectives on writing. We would not suggest separate instruction for different genders, but we do think that it is important to be aware of these differences and to keep them in mind for classroom practice. When guiding students into learning how to write, it may be beneficial to offer support depending on students' needs. For example, as girls' self-efficacy is lower on many aspects than boys', it could help them to reflect on their strong points and to formulate, with the guide of the teacher, a plan of action to deal with the aspects they feel less confident about. Boys may need more guidance from the teacher to help them figure out a writing style, as they do not tend to have a specific preference for certain writing styles such as pre-planning or post-draft revision. Thirdly, as we found an effect of genre on synthesis writing, future writing process instruction should take this

into account. For example, when designing videos that model writing approaches, the models could differ depending on the genre of the task: in the case of an informative synthesis, students switch between the sources in the first interval, indicating that they compare and contrast the sources; in the case of the argumentative synthesis, students switch between the synthesis and the sources, indicating that they already select information from the sources to include in their text.

The methodology was of utmost importance for this descriptive study to form a baseline for future studies on synthesis writing. We would like to point out the distinctive characteristics of this national survey that contribute to making it fit as a baseline study, and in that way open up possibilities for future research building on this work. First, it contains a large and representative sample. Great care was given to careful sampling as to obtain a sample that is representative for the population of Dutch students in the three highest grades of upper-secondary education. Secondly, for generalisation purposes, students performed several tasks and a wide range of synthesis tasks was created. Most of the students wrote four synthesis tasks as more than one task is needed to get a valid and reliable view of a student's writing performance (Van den Bergh et al., 2012). For each of the four tasks, different variants were created. The variety of tasks - varying in genre, relation between sources, and amount of irrelevant information in the sources - allows us to generalise the findings. Thirdly, our sample contains a variety of data: product, process and perspectives. Attention was given to each of these three aspects as to make the data as rich and reliable as possible. Text quality was assessed by a panel of overlapping rater teams, who all received a training. Rating scales with benchmark texts were used, a highly reliable rating method (Van Steendam, 2017). For the writing processes, the keystroke logging program Inputlog was used. The logfiles were carefully filtered and inspected on errors. Following Breetvelt, Van den Bergh, and Rijlaarsdam (1994) who stress the importance of time allocation of different

writing process aspects, both global and interval-based aspects were investigated. And regarding perspectives on writing, we built on existing questionnaires and checked the validity on our data sample.

Regarding future research and limitations of this study, there is the possibility of expanding this national survey sample with data on bachelor students as to further map the development of synthesis writing. Another possibility is to add more data of grade 12 students as this group is somewhat underrepresented in the current sample. Moreover, based on future research and developments in keystroke logging, other analyses (for example, more detailed revision analyses) could be done with regard to the writing process variables.

In conclusion, by describing the text quality and writing process of 2310 synthesis texts, we know now what students can accomplish and how they write (two genres of) source-based texts in three grades of upper-secondary education. Besides a clear view on the current state of affairs, it also allowed us to map the development of synthesis writing over the grades (De Glopper, 1988). Though the information obtained from this study is purely descriptive, we believe it has a great value: it forms a baseline for further research with the goal to improve synthesis writing in secondary education. Based on this national survey study, we know how students perform, how they write a synthesis text, and what their perspectives on writing are. Moreover, we know how these aspects change over the grades, and what the effect of genre and gender are. In this way, this study gives us insight in aspects to focus on for further research. For example, we used this national survey as a baseline for two studies on the relation between the writing process and text quality for two genres of synthesis texts (Van Steendam et al., 2020; Vandermeulen, Van Steendam, Van den Broek, et al., 2020). The selection of the process variables for these process-product studies was based on the national survey results: we mainly used variables that changed over the grades and on which genre had an effect. Moreover, we also conducted an

intervention study in which we gave feedback and instruction on the writing process (Vandermeulen, Van Steendam, & Rijlaarsdam, 2020). The results from the national baseline study provided input for the selection of process variables to be included in this intervention.

Acknowledgements

We would like to thank prof. dr. Luuk Van Waes and prof. dr. Mariëlle Leijten (University of Antwerp) for their valuable contributions and advice in organising a large data collection with Inputlog; and Tom Pauwaert and Eric Van Horenbeek for creating a user-friendly database. We would also like to thank Brenda van den Broek, former PhD student on the project, and the team of student-assistants for their help in collecting the data. Special thanks to all the participating schools, teachers and students; and all the people that contributed to rating the texts.

References

Anmarkrud, Ø., Bråten, I., & Strømsø, H. I. (2014). Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents. *Learning and Individual Differences*, 30, 64–76. <https://doi.org/10.1016/j.lindif.2013.01.007>

Barzilai, S., Zohar, A. R., & Mor-Hagani, S. (2018). Promoting Integration of Multiple Texts: a Review of Instructional Approaches and Practices. *Educational Psychology Review*. <https://doi.org/10.1007/s10648-018-9436-8>

Bazerman, C. (1994). Systems of Genres and the Enactment of Social Intentions. In A. Freedman & P. Medway (Eds.), *Genre and the New Rhetoric* (pp. 67–85). Taylor & Francis.

Blok, H. (1986). Essay Rating by the Comparison Method. *Tijdschrift Voor Onderwijsresearch*, 11(4), 169–176.

Boscolo, P., Arfé, B., & Quarisa, M. (2007). Improving the quality of students' academic writing: an intervention study. *Studies in Higher Education*, 32(4), 419–438. <https://doi.org/10.1080/03075070701476092>

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, 32(1), 83–100. <https://doi.org/10.1177/0265532214542994>

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. In *Bringing writing research into the classroom. The effectiveness of Tekster, a newly developed writing program for elementary students (Doctoral dissertation)* (pp. 63–82). University of Utrecht.

Bouwer, R., Koster, M., & Van den Bergh, H. (2018). Effects of a strategy-focused instructional program on the writing quality of upper elementary students in the Netherlands. *Journal of Educational Psychology*, 110(1), 58–71. <https://doi.org/10.1037/edu0000206>

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

Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: when and how? *Cognition and Instruction*, 12(2), 103–123. https://doi.org/10.1207/s1532690xc1202_2

Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1996). Rereading and generating and their relation to text quality. An application of multilevel analysis on writing process data. In G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 10–20). Amsterdam University Press. <https://doi.org/https://doi.org/10.5117/9789053561973>

Cordeiro, C., Castro, S. L., & Limpo, T. (2018). Examining Potential Sources of Gender Differences in Writing: The Role of Handwriting Fluency and Self-Efficacy Beliefs. *Written Communication*, 35(4), 448–473. <https://doi.org/10.1177/0741088318788843>

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognitive Development*, 11(2), 121–136. <https://doi.org/10.1080/15248371003699969>

De Glopper, K. (1988). *Schrijven beschreven. Inhoud, opbrengsten en achtergronden van het schrijfonderwijs in de eerste vier leerjaren van*

- het voortgezet onderwijs. SVO. <http://taalunie-versum.org/onderwijs/onderzoek/1969-1997/artikel.php?h=4.5>
- De Smedt, F., Van Keer, H., & Merchie, E. (2016). Student, teacher and class-level correlates of Flemish late elementary school children's writing performance. *Reading and Writing*, 29(5), 833–868. <https://doi.org/10.1007/s11145-015-9590-z>
- Drijbooms, E. (2016). *Cognitive and Linguistic Factors in Writing Development*. Radboud University Nijmegen.
- Du, H., & List, A. (2020). Researching and writing based on multiple texts. *Learning and Instruction*, 66. <https://doi.org/10.1016/j.learninstruc.2019.101297>
- Expertgroep Doorlopende Leerlijnen. (2009). *Referentiekader taal en rekenen. De referentieniveaus [Frame of reference language and arithmetic]*.
- Feddema, M., & Hoek, P. (2018). Lezen en schrijven hand in hand in synthesesetaken. *Levende Talen Magazine*, 8, 18–23.
- Graham, S. (2018). A Revised Writer(s)-Within-Community Model of Writing. *Educational Psychologist*, 53(4), 258–279. <https://doi.org/10.1080/00461520.2018.1481406>
- Hox, J. J. (2002). Multilevel analysis: Techniques and applications: Second edition. In *Multilevel Analysis: Techniques and Applications*. Erlbaum. <https://doi.org/10.4324/9780203852279>
- Inspectie van het Onderwijs. (2010). *Het onderwijs in het schrijven van teksten*.
- Kieft, M., Rijlaarsdam, G., Galbraith, D., & Van den Bergh, H. (2007). The effects of adapting a writing course to students' writing strategies. *The British Journal of Educational Psychology*, 77, 565–578. <https://doi.org/10.1348/096317906X120231>
- Knospe, Y. (2017). *Writing in a third language. A study of upper-secondary students' texts, writing processes and metacognition*. (Doctoral dissertation) Umeå University.
- Kuhlemeier, H., & Van den Bergh, H. (1990). Peilingsonderzoek in het voortgezet onderwijs: de proefpeiling Nederlands. *Levende Talen Magazine*, 77(447), 24–29.
- Kuhlemeier, H., Van Til, A., & Van den Bergh, H. (2014). Schrijfvaardigheid Nederlands vergeleken met de referentieniveaus: een verkenning. *Levende Talen Tijdschrift*, 15(2), 37–46.
- Leijten, M., & Van Waes, L. (2013). Keystroke Logging in Writing Research: Using Inputlog to Analyze and Visualize Writing Processes. *Written Communication*, 30(3), 358–392. <https://doi.org/10.1177/0741088313491692>
- Lesterhuis, M., Verhavert, S., Coertjens, L., Donche, V., & De Maeyer, S. (2016). Comparative Judgment as a promising alternative to score competences. In E. Cano & G. Ion (Eds.), *Innovative Practices for Higher Education Assessment and Measurement* (pp. 119–138). IGI Global. <https://doi.org/10.4018/978-1-5225-0531-0>
- Limpo, T., & Alves, R. A. (2017). Relating beliefs in writing skill malleability to writing performance: The mediating role of achievement goals and self-efficacy. *Journal of Writing Research*, 9(2), 97–125. <https://doi.org/10.17239/jowr-2017.09.02.01>
- Martínez, I., Mateos, M., Martín, E., & Rijlaarsdam, G. (2015). Learning history by composing synthesis texts: Effects of an instructional programme on learning, reading and writing processes, and text quality. *Journal of Writing Research*, 7(2), 275–302. <https://doi.org/10.17239/jowr-2015.07.02.03>
- Mateos, M., Martín, E., Villalón, R., & Luna, M. (2008). Reading and writing to learn in secondary education: Online processing activity and written products in summarizing and synthesizing tasks. *Reading and Writing*, 21(7), 675–697. <https://doi.org/10.1007/s11145-007-9086-6>
- Mateos, M., & Solé, I. (2009). Synthesizing information from various texts: A study of procedures and products at different educational levels. *European Journal of Psychology of Education*, 24(4), 435–451. <https://doi.org/10.1007/BF03178760>
- National Center for Education Statistics. (2012). *The Nation's Report card: Writing 2011*.
- Olive, T., & Kellogg, R. T. (2002). Concurrent activation of high- and low-level production processes in written composition. *Memory and Cognition*, 30(4), 594–600. <https://doi.org/10.3758/BF03194960>
- Pollitt, A. (2012). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2), 157–170. <https://doi.org/10.1007/s10798-011-9189-x>
- Pollmann, E., Prenger, J., & de Glopper, K. (2012).

- Het beoordelen van leerlingteksten met behulp van een schaalmodel. *Levende Talen Tijdschrift*, 13(3), 15–24.
- Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, 59(4), 413–425. <https://doi.org/10.1016/j.jml.2008.02.002>
- Rietdijk, S., Janssen, T., Van Weijen, D., Van den Bergh, H., & Rijlaarsdam, G. (2017). Improving writing in primary schools through a comprehensive writing program. *Journal of Writing Research*, 9(2), 173–225. <https://doi.org/10.17239/jowr-2017.09.02.04>
- Rijlaarsdam, G. (1986). *Effecten van leerlingenrespons op aspecten van stelvaardigheid [Effects of student peer feedback on some aspects of written composition skills]*. (Doctoral dissertation) Foundation Center for Educational Research Report 88, University of Amsterdam.
- Rijlaarsdam, G., & Schoonen, R. (1988). Effects of a teaching program based on peer evaluation on written composition and some variables related to writing apprehension. In *SCO Cahier* (Vol. 47). <https://eric.ed.gov/?id=ED311469>
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*. <https://doi.org/10.1191/0265532205lt295oa>
- Sijtsma, J. (1997). Balans van het taalonderwijs aan het einde van de basisschool 2. Uitkomsten van de tweede taalpeiling einde basisonderwijs. In *PPON-reeks* (Vol. 10). Cito Instituut voor Toetsontwikkeling. <http://taazoe.live.statik.be/nodes/publicatiedetail/nl/balans-van-het-taalonderwijs-aan-het-einde-van-de-basisschool-2>
- Sijtsma, J., van de Schoot, F., & Hemker, B. (1998). Balans van het taalonderwijs aan het einde van de basisschool 3. Uitkomsten van de derde peiling in 1998. In *PPON-reeks* (Vol. 19). Cito. <http://taazoe.live.statik.be/nodes/publicatiedetail/nl/balans-van-het-taalonderwijs-aan-het-einde-van-de-basisschool-3>
- Solé, I., Miras, M., Castells, N., Espino, S., & Minguella, M. (2013). Integrating Information: An Analysis of the Processes Involved and the Products Generated in a Written Synthesis Task. *Written Communication*, 30(1), 63–90. <https://doi.org/10.1177/0741088312466532>
- Spivey, N., & King, J. (1989). Readers as writers composing from sources. *Reading Research Quarterly*, 24(1), 7–26. <https://doi.org/10.2307/748008>
- Swales, J. (1990). Genre Analysis: English in Academic and Research Settings. In *Language*. Cambridge University Press. <https://doi.org/10.2307/416471>
- Tillema, M., Van den Bergh, H., Rijlaarsdam, G., & Sanders, T. (2013). Quantifying the quality difference between L1 and L2 essays: A rating procedure with bilingual raters and L1 and L2 benchmark essays. *Language Testing*. <https://doi.org/10.1177/0265532212442647>
- Van den Bergh, H., De Maeyer, S., 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 Practices* (pp. 23–32). Brill. [https://doi.org/10.1108/S1572-6304\(2012\)0000027005](https://doi.org/10.1108/S1572-6304(2012)0000027005)
- Van den Bergh, H., & Eiting, M. H. (1989). A method of estimating rater reliability. *Journal of Educational Measurement*, 26(1), 29–40.
- Van den Bergh, H., & Rijlaarsdam, G. (2001). Changes in cognitive activities during the writing process and relationships with text quality. *Educational Psychology*, 21(4), 373–385. <https://doi.org/10.1080/01443410120090777>
- Van Ockenburg, L., Van Weijen, D., & Rijlaarsdam, G. (2018). Syntheseteksten leren schrijven in het voortgezet onderwijs. Het verband tussen schrijfaanpak en voorkeur voor leeractiviteiten. *Levende Talen Tijdschrift*, 19(2), 3–14.
- Van Steendam, E. (2017). Een synopsis van schrijfonderwijsonderzoek in Nederland en Vlaanderen: waar staan we en waar willen we naartoe? *Pedagogische Studiën*, 94(4), 348–359.
- Van Steendam, E., & Bouwer, R. (2018). Measuring writing. Defining and operationalizing the construct of writing quality. *Sig Writing Research School*.
- Van Steendam, E., Vandermeulen, N., De Maeyer, S., & Rijlaarsdam, G. (2020). Dynamic writing process configurations in synthesis tasks. *[In Preparation]*.
- Vandermeulen, N., Van Steendam, E., & Rijlaarsdam, G. (2020). The effect of process-oriented feedback on source-based writing: A keystroke logging study. *[In Preparation]*. <https://doi.org/10.1007/s11145-019-09958-3>

- Villalón, R., Mateos, M., & Cuevas, I. (2015). High school boys' and girls' writing conceptions and writing self-efficacy beliefs: what is their role in writing performance? *Educational Psychology*, 35(6), 653–674. <https://doi.org/10.1080/01443410.2013.836157>
- White, M. J., & Bruning, R. (2005). Implicit writing beliefs and their relation to writing quality. *Contemporary Educational Psychology*, 30(2), 166–189. <https://doi.org/10.1016/j.cedpsych.2004.07.002>
- Zwarts, M. (1990). Balans van het taalonderwijs aan het einde van de basisschool. Uitkomsten van de eerste taalpeiling einde basisonderwijs. In *PPPON-reeks* (Vol. 2). Cito Instituut voor Toetsontwikkeling. <http://taazoe.live.statik.be/nodes/publicatiedetail/nl/balans-van-het-taalonderwijs-aan-het-einde-van-de-basisschool>

Authors

As a member of the LIFT project team, **Nina Vandermeulen** obtained her PhD on synthesis writing at the University of Antwerp and KU Leuven. **Sven De Maeyer** is professor at the Training and Education Sciences department of the University of Antwerp. **Elke Van Steendam** is assistant professor at the Language and Education department at KU Leuven. **Marije Lesterhuis** worked as a post-doctoral researcher on the LIFT project at the University of Antwerp. **Huib van den Bergh** is professor at the Language, Literature, and Communication department at the University of Utrecht. **Gert Rijlaarsdam** is professor at the Graduate School of Child Development and Education at the University of Amsterdam.

Correspondence: Nina Vandermeulen, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Belgium nina.vandermeulen@uantwerpen.be

Het schrijven van een synthesesetext - een tekst waarin informatie uit verschillende bronnen geïntegreerd wordt - maakt deel uit van het curriculum in het Nederlandse vwo-onderwijs. Deze studie bestaat uit een nationale peiling naar de synthesevaardigheid in de drie hoogste leerjaren van het vwo. Het doel van deze studie was om drie aspecten van syntheseschrijven in kaart te brengen: tekstkwaliteit, schrijfproces en leerlingperspectief op schrijven. Een representatieve steekproef van 658 leerlingen nam deel; elke leerling schreef vier teksten. Teksten werden beoordeeld met behulp van tekstschalen met benchmarks; het schrijfproces werd geregistreerd met keystroke logging; en de perspectieven van de leerlingen op schrijven werden gemeten met een vragenlijst. Via multilevel analyses gingen we het effect van leerjaar, geslacht en tekstgenre (argumentatieve/informatieve synthese) na op tekstkwaliteit en schrijfproces, en het effect van leerjaar en geslacht op de perspectieven. Deze nationale peiling is een beschrijvend onderzoek dat inzicht biedt in de huidige stand van zaken omtrent syntheseschrijven: hoe presteren leerlingen op synthesesetaken?, hoe schrijven ze syntheseseteksten?, en wat zijn hun perspectieven op het schrijven van een synthesesetext? Bovendien dient deze studie ook als baseline voor toekomstig onderzoek.

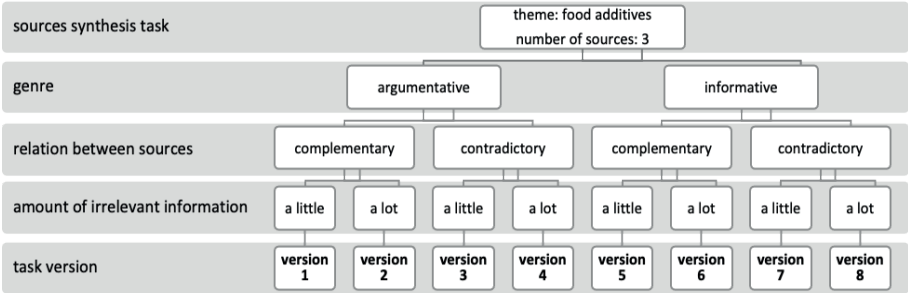
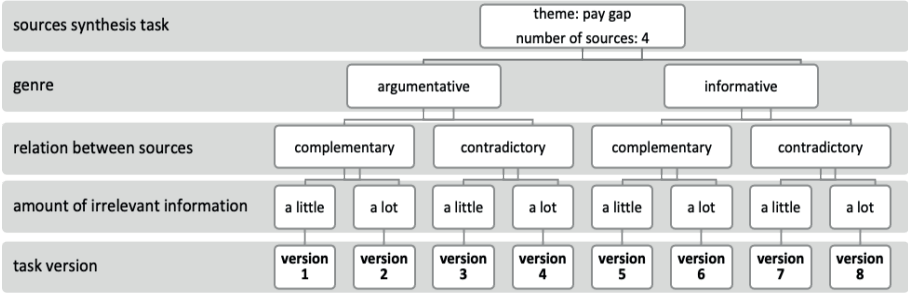
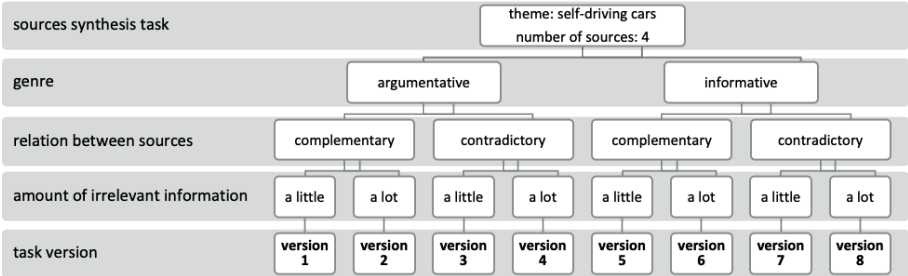
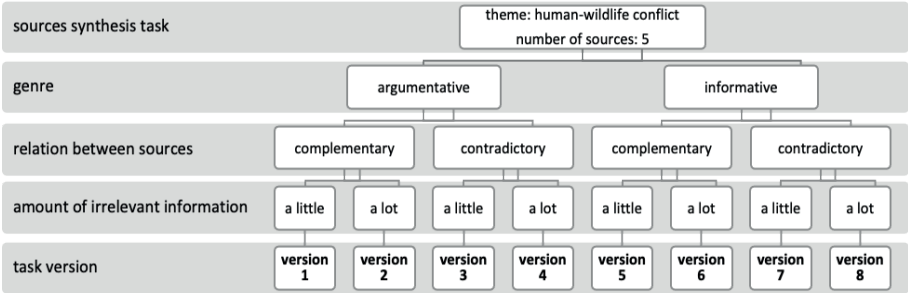
Keywords: Nationale peiling, syntheseschrijven, keystroke logging, schrijfproces, schrijfonderwijs

Samenvatting

Syntheseschrijven in het hoger secundair onderwijs in Nederland in kaart gebracht
Een nationale peiling naar tekstkwaliteit, schrijfproces en schrijverskenmerken

Appendices

Appendix A. Visualisation of the task construction



Appendix B. Task instructions for the participants

Type of instruction	Clarification
Explanation on what a synthesis text is	<p>A synthesis is a text based on various sources. Your text brings together the information from the different sources. When reading a synthesis you should be able to understand the text without having read the sources.</p> <p>How do you write a synthesis?</p> <ul style="list-style-type: none">- You start by reading the sources- You select the information you need, to write a new text about theme X.- You bring together the information from the different sources and connect the sources. In this way you integrate the information from the sources into a new independent text.- You elaborate your synthesis by writing a text that is understandable for people who have not read the sources.
Explanation on the characteristics of an argumentative/ informative synthesis text	<ul style="list-style-type: none">- Informative synthesis: Your text gives a concise and at the same time clear overview of the situation. You describe the situation concerning theme X in a neutral manner, that is, without taking position.- Argumentative synthesis: In your text you defend the following point of view: X. You support this point of view with arguments from the source texts.
Instructions on how to deal with the sources	Use only the relevant information and use information from all offered source texts.
Instructions on the target audience	Your text has to be understandable to peers who did not read the source texts.
Instructions on style	Use your own words, avoid copying from the sources.
Instructions on text length	Write a text of approximately 350 words
Time indication	You have 50 minutes to read the sources and to write your text. Divide your time between reading and writing. Write the best possible text in this given time.

Appendix C. Overview of the scores for information, integration, coherence and cohesion, language and global judgment for the selected benchmark texts

Benchmark	Information score	Integration score	Cohesion/ Coherence score	Language score	Global score
Argumentative Genre					
50 (-2 SD)	-1,32	-0,40	-1,34	-1,96	-1,89
75 (-1 SD)	-0,63	-1,33	-0,70	1,17	-0,87
100 (average)	0,09	-0,08	0,28	-0,60	-0,15
125 (+1 SD)	0,86	-1,28	0,89	0,33	1,02
150 (+2 SD)	- *	-	-	-	-
Informative Genre					
50 (-2 SD)	-1,32	-1,64	-1,223	-2,24	-2,48
75 (-1 SD)	0,89	-1,19	-1,85	0,19	-1,01
100 (average)	0,34	-0,66	-0,69	-0,86	0,13
125 (+1 SD)	1,21	1,48	1,52	0,51	1,18
150 (+2 SD)	1,52	1,33	3,28	2,81	1,70

* Note: The 150 benchmark text for the argumentative genre was added after the assessment in D-PAC, thus we do not dispose of the individual scores for this text. Reason for this was the absence of a representative benchmark in the subsample.

Appendix E. Linear mixed models - Model fit and comparison

Text quality					
	Model Fit -2LL	Model Comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	19563				
Model 1 (gender, grade, genre)	19473	1 vs 0	90.672	4	< .001
Model 2 (grade x gender)	19469	2 vs 1	3.6025	2	.017
Model 3 (grade x genre)	19460	3 vs 1	13.011	2	.001
Total process time					
	Model Fit -2LL	Model Comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5293.2				
Model 1 (gender, grade, genre)	5256.6	1 vs 0	36.635	4	<.001
Model 2 (grade x gender)	5256.1	2 vs 1	.453	2	.797
Model 3 (grade x genre)	5244.9	3 vs 1	11.681	2	.003
Proportion of time in sources - Interval 1					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5529				
Model 1 (gender, grade, genre)	5503.4	1 vs 0	25.535	4	<.001
Model 2 (grade x gender)	5501.5	2 vs 1	1.8914	2	.388
Model 3 (grade x genre)	5502.8	3 vs 1	.6067	2	.738
Proportion of time in sources - Interval 2					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5646.1				
Model 1 (gender, grade, genre)	5623.5	1 vs 0	22.581	4	<.001
Model 2 (grade x gender)	5612.5	2 vs 1	11.006	2	.004
Model 3 (grade x genre)	5622.8	3 vs 1	.6497	2	.723
Proportion of time in sources - Interval 3					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5729.3				
Model 1 (gender, grade, genre)	5678.4	1 vs 0	50.908	4	<.001
Model 2 (grade x gender)	5673.3	2 vs 1	5.1185	2	.077
Model 3 (grade x genre)	5674.1	3 vs 1	4.3372	2	.114

Proportion of active writing time - Interval 1					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5363.9				
Model 1 (gender, grade, genre)	5334.3	1 vs 0	29.642	4	<.001
Model 2 (grade x gender)	5333.5	2 vs 1	.7901	2	.674
Model 3 (grade x genre)	5333.4	3 vs 1	.8645	2	.649
Proportion of active writing time - Interval 2					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5513.2				
Model 1 (gender, grade, genre)	5503.8	1 vs 0	9.4042	4	.052
Model 2 (grade x gender)	5501.5	2 vs 1	2.3574	2	.308
Model 3 (grade x genre)	5501	3 vs 1	2.7982	2	.247
Proportion of active writing time - Interval 3					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5525.2				
Model 1 (gender, grade, genre)	5518.8	1 vs 0	6.3431	4	.175
Model 2 (grade x gender)	5516	2 vs 1	2.8546	2	.240
Model 3 (grade x genre)	5516.9	3 vs 1	1.9294	2	.381
Proportion of pause time during production - Interval 1					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5509.4				
Model 1 (gender, grade, genre)	5499.9	1 vs 0	9.5091	4	.050
Model 2 (grade x gender)	5496.9	2 vs 1	3.0047	2	.223
Model 3 (grade x genre)	5495.6	3 vs 1	4.3347	2	.115
Proportion of pause time during production - Interval 2					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5350.5				
Model 1 (gender, grade, genre)	5325.7	1 vs 0	24.777	4	<.001
Model 2 (grade x gender)	5322	2 vs 1	3.7604	2	.153
Model 3 (grade x genre)	5324.2	3 vs 1	1.4959	2	.473

Proportion of pause time during production - Interval 3					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5557.8				
Model 1 (gender, grade, genre)	5540.3	1 vs 0	17.544	4	.002
Model 2 (grade x gender)	5540.2	2 vs 1	.1054	2	.949
Model 3 (grade x genre)	5537.4	3 vs 1	2.8788	2	.237
Total number of keystrokes typed					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5026.7				
Model 1 (gender, grade, genre)	4950.9	1 vs 0	75.766	4	<.001
Model 2 (grade x gender)	4949.1	2 vs 1	1.7757	2	.412
Model 3 (grade x genre)	4946.9	3 vs 1	4.0084	2	.135
Number of keystrokes per minute - Interval 1					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5088				
Model 1 (gender, grade, genre)	5066.3	1 vs 0	21.782	4	<.001
Model 2 (grade x gender)	5064.2	2 vs 1	2.0793	2	.354
Model 3 (grade x genre)	5065.8	3 vs 1	.4605	2	.7943
Number of keystrokes per minute - Interval 2					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5261.7				
Model 1 (gender, grade, genre)	5242.7	1 vs 0	18.994	4	<.001
Model 2 (grade x gender)	5240.6	2 vs 1	2.0911	2	.352
Model 3 (grade x genre)	5240.3	3 vs 1	2.3767	2	.305
Number of keystrokes per minute - Interval 3					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5477				
Model 1 (gender, grade, genre)	5447.6	1 vs 0	29.39	4	<.001
Model 2 (grade x gender)	5446.7	2 vs 1	.9213	2	.631
Model 3 (grade x genre)	5441.1	3 vs 1	6.4864	2	.039

Number of pauses per minute- Interval 1					
	Model fit -2LL		Model comparison χ^2_{change}	df_{change}	p
Model 0 (intercept)	5686.3				
Model 1 (gender, grade, genre)	5678.1	1 vs 0	8.2035	4	.084
Model 2 (grade x gender)	5675.7	2 vs 1	2.4214	2	.298
Model 3 (grade x genre)	5675.6	3 vs 1	2.5449	2	.280
Number of pauses per minute- Interval 2					
	Model fit -2LL		Model comparison χ^2_{change}	df_{change}	p
Model 0 (intercept)	5533.3				
Model 1 (gender, grade, genre)	5520.5	1 vs 0	12.826	4	.012
Model 2 (grade x gender)	5513.7	2 vs 1	6.7619	2	.034
Model 3 (grade x genre)	5519.9	3 vs 1	.5862	2	.746
Number of pauses per minute- Interval 3					
	Model fit -2LL		Model comparison χ^2_{change}	df_{change}	p
Model 0 (intercept)	5709.8				
Model 1 (gender, grade, genre)	5670.6	1 vs 0	39.172	4	<.001
Model 2 (grade x gender)	5668.9	2 vs 1	1.6902	2	.430
Model 3 (grade x genre)	5662.4	3 vs 1	8.2328	2	.016
Mean pause time- Interval 1					
	Model fit -2LL		Model comparison χ^2_{change}	df_{change}	p
Model 0 (intercept)	5981.7				
Model 1 (gender, grade, genre)	5968.3	1 vs 0	13.456	4	.009
Model 2 (grade x gender)	5967.1	2 vs 1	1.1615	2	.560
Model 3 (grade x genre)	5963.4	3 vs 1	4.8585	2	.088
Mean pause time- Interval 2					
	Model fit -2LL		Model comparison χ^2_{change}	df_{change}	p
Model 0 (intercept)	5790.4				
Model 1 (gender, grade, genre)	5776.8	1 vs 0	13.647	4	.009
Model 2 (grade x gender)	5776.7	2 vs 1	.0483	2	.976
Model 3 (grade x genre)	5775.9	3 vs 1	.9101	2	.634

Mean pause time- Interval 3					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5797.6				
Model 1 (gender, grade, genre)	5795.8	1 vs 0	1.8395	4	.765
Model 2 (grade x gender)	5793.4	2 vs 1	2.3861	2	.303
Model 3 (grade x genre)	5794	3 vs 1	1.7928	2	.408
Produced ratio					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	5218.4				
Model 1 (gender, grade, genre)	5114.1	1 vs 0	104.32	4	<.001
Model 2 (grade x gender)	5114	2 vs 1	.0845	2	.959
Model 3 (grade x genre)	5104.5	3 vs 1	9.5984	2	.008
Number of transitions per minute between the sources - Interval 1					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	4941.9				
Model 1 (gender, grade, genre)	4932.4	1 vs 0	9.4736	4	.050
Model 2 (grade x gender)	4931.2	2 vs 1	1.2507	2	.535
Model 3 (grade x genre)	4932.3	3 vs 1	.1072	2	.948
Number of transitions per minute between the sources - Interval 2					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	3541.4				
Model 1 (gender, grade, genre)	3530.3	1 vs 0	11.078	4	.026
Model 2 (grade x gender)	3522.9	2 vs 1	7.442	2	.024
Model 3 (grade x genre)	3529.8	3 vs 1	.5301	2	.767
Number of transitions per minute between the sources - Interval 3					
	Model fit -2LL	Model comparison			
			χ^2_{change}	df_{change}	p
Model 0 (intercept)	3460.8				
Model 1 (gender, grade, genre)	3454	1 vs 0	6.8257	4	.145
Model 2 (grade x gender)	3442.6	2 vs 1	11.406	2	.003
Model 3 (grade x genre)	3453.3	3 vs 1	.7293	2	.694

Number of transitions per minute between the synthesis text and the sources - Interval 1					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	3955.1				
Model 1 (gender, grade, genre)	3898.8	1 vs 0	56.2688	4	< .001
Model 2 (grade x gender)	3896.5	2 vs 1	2.2835	2	.319
Model 3 (grade x genre)	3897.6	3 vs 1	1.2309	2	.541
Number of transitions per minute between the synthesis text and the sources - Interval 2					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	4225.1				
Model 1 (gender, grade, genre)	4207.8	1 vs 0	17.3738	4	.002
Model 2 (grade x gender)	4204.6	2 vs 1	3.1604	2	.206
Model 3 (grade x genre)	4205.2	3 vs 1	2.5766	2	.276
Number of transitions per minute between the synthesis text and the sources - Interval 3					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	3335.6				
Model 1 (gender, grade, genre)	3284	1 vs 0	51.569	4	< .001
Model 2 (grade x gender)	3283.2	2 vs 1	.784	2	.676
Model 3 (grade x genre)	3283.3	3 vs 1	.7364	2	.692
Writing apprehension - Cognitive					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1841.3				
Model 1 (grade, gender)	1845.2	1 vs 0	8.098	3	.044
Model 2 (grade x gender)	1849.2	2 vs 1	0.06	2	.971
Writing apprehension - Affective					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1813.4				
Model 1 (grade, gender)	1798.9	1 vs 0	14.463	3	.002
Model 2 (grade x gender)	1797.3	2 vs 1	1.571	2	.456

Writing apprehension - Evaluative					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1847.1				
Model 1 (grade, gender)	1840.2	1 vs 0	9.928	3	.074
Model 2 (grade x gender)	1836.7	2 vs 1	3.484	2	.175
Writing beliefs - Transmission					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1835.8				
Model 1 (grade, gender)	1827.7	1 vs 0	8.069	3	.045
Model 2 (grade x gender)	1827.2	2 vs 1	0.524	2	.769
Writing beliefs - Emotional engagement					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1843.9				
Model 1 (grade, gender)	1834.5	1 vs 0	9.443	3	.024
Model 2 (grade x gender)	1830.9	2 vs 1	3.591	2	.166
Writing beliefs - High amount of revision					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1835.5				
Model 1 (grade, gender)	1824.8	1 vs 0	10.75	3	.013
Model 2 (grade x gender)	1822.1	2 vs 1	2.678	2	.262
Writing beliefs - Cognitive engagement					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1840.8				
Model 1 (grade, gender)	1828.1	1 vs 0	12.716	3	.005
Model 2 (grade x gender)	1827.1	2 vs 1	1.054	2	.590
Self-efficacy - Dealing with sources					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1824.9				
Model 1 (grade, gender)	1816.3	1 vs 0	8.651	3	.034
Model 2 (grade x gender)	1815.4	2 vs 1	0.893	2	.640

Self-efficacy - Language use					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1831.7				
Model 1 (grade, gender)	1819.9	1 vs 0	11.736	3	.008
Model 2 (grade x gender)	1817.4	2 vs 1	2.501	2	.286
Self-efficacy - Concise writing					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1818				
Model 1 (grade, gender)	1813.8	1 vs 0	4.259	3	.235
Model 2 (grade x gender)	1808.9	2 vs 1	4.88	2	.087
Self-efficacy - Text structure					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1834				
Model 1 (grade, gender)	1823.8	1 vs 0	10.243	3	.017
Model 2 (grade x gender)	1817.6	2 vs 1	6.195	2	.045
Self-efficacy - Integration of sources					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1825.4				
Model 1 (grade, gender)	1818.5	1 vs 0	6.860	3	.076
Model 2 (grade x gender)	1814.6	2 vs 1	3.915	2	.141
Self-efficacy - Elaboration of sources					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1833.9				
Model 1 (grade, gender)	1826.1	1 vs 0	7.806	3	.050
Model 2 (grade x gender)	1821.8	2 vs 1	4.324	2	.115
Writing style - Preplanning					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1828.5				
Model 1 (grade, gender)	1814.2	1 vs 0	14.384	3	.002
Model 2 (grade x gender)	1812.8	2 vs 1	1.342	2	.511

Writing style - Post-draft revision					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1837.6				
Model 1 (grade, gender)	1824	1 vs 0	13.666	3	.003
Model 2 (grade x gender)	1820	2 vs 1	3.9	2	.142
Writing style - Short production cycles					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1843.6				
Model 1 (grade, gender)	1832.1	1 vs 0	11.549	3	.009
Model 2 (grade x gender)	1831.8	2 vs 1	0.28	2	.869
Writing style - Difficult idea generation					
	Model fit	Model comparison			
	-2LL		χ^2_{change}	df_{change}	p
Model 0 (intercept)	1846.5				
Model 1 (grade, gender)	1843.3	1 vs 0	3.148	3	.369
Model 2 (grade x gender)	1842.8	2 vs 1	0.508	2	.776

Appendix F. Parameter estimates for the best fitting models

* Reference category intercept: Grade 11 - Genre ARG - Gender Female

Mean Score Text Quality			
Predictors	Estimates	CI	p
(Intercept)	94.56	91.88 – 97.25	<0.001
Grade 10	-10.37	-12.85 – -7.88	<0.001
Grade 12	4.35	0.75 – 7.95	0.019
Gender Male	-3.98	-6.07 – -1.89	<0.001
Genre INF	-1.87	-4.66 – 0.92	0.194
Grade10:GenreINF	4.70	2.15 – 7.25	<0.001
Grade12:GenreINF	2.16	-1.16 – 5.49	0.202
Random Effects			
σ^2	197.02		
τ_{00} Participant	93.13		
τ_{00} School	11.07		
τ_{00} Task	9.59		
ICC _{Participant}	0.30		
ICC _{School}	0.04		
ICC _{Task}	0.03		
Observations	2310		
Marginal R ² / Conditional R ²	0.089 / 0.422		
Total process time			
Predictors	Estimates	CI	p
(Intercept)	0.19	0.01 – 0.37	0.040
Grade 10	-0.21	-0.36 – -0.06	0.005
Grade 12	0.19	-0.03 – 0.40	0.090
Gender Male	-0.36	-0.49 – -0.23	<0.001
Genre INF	-0.06	-0.25 – 0.12	0.509
Grade10:GenreINF	0.21	0.09 – 0.34	0.001
Grade12:GenreINF	0.08	-0.08 – 0.24	0.355
Random Effects			
σ^2	0.41		
τ_{00} Participant	0.43		
τ_{00} School	0.05		
τ_{00} Task	0.06		
ICC _{Participant}	0.45		
ICC _{School}	0.05		
ICC _{Task}	0.06		
Observations	2155		
Marginal R ² / Conditional R ²	0.046 / 0.584		
Proportion of time in sources - Interval 1			
Predictors	Estimates	CI	p
(Intercept)	-0.15	-0.31 – 0.01	0.079
Grade 10	0.12	-0.01 – 0.26	0.070
Grade 12	-0.25	-0.43 – -0.08	0.005
Gender Male	0.14	0.01 – 0.27	0.035
Genre INF	0.20	0.02 – 0.37	0.034

Random Effects			
σ^2	0.49		
T_{00} Participant	0.43		
T_{00} School	0.01		
T_{00} Task	0.06		
ICC Participant	0.43		
ICC School	0.01		
ICC Task	0.06		
Observations	2155		
Marginal R^2 / Conditional R^2	0.031 / 0.512		
Proportion of time in sources - Interval 2			
Predictors	Estimates	CI	p
(Intercept)	-0.15	-0.31 – 0.00	0.055
Grade 10	0.20	0.04 – 0.36	0.015
Grade 12	-0.30	-0.52 – -0.08	0.007
Gender Male	0.10	-0.08 – 0.28	0.286
Genre INF	0.24	0.11 – 0.37	0.001
Grade10:GenderMale	-0.34	-0.61 – -0.07	0.013
Grade12:GenderMale	0.22	-0.14 – 0.57	0.231
Random Effects			
σ^2	0.55		
T_{00} Participant	0.35		
T_{00} School	0.03		
T_{00} Task	0.03		
ICC Participant	0.37		
ICC School	0.03		
ICC Task	0.03		
Observations	2155		
Marginal R^2 / Conditional R^2	0.037 / 0.447		
Proportion of time in sources - Interval 3			
Predictors	Estimates	CI	p
(Intercept)	-0.18	-0.30 – -0.06	0.005
Grade 10	0.16	0.03 – 0.29	0.013
Grade 12	-0.30	-0.47 – -0.12	0.001
Gender Male	0.06	-0.06 – 0.18	0.324
Genre INF	0.30	0.21 – 0.39	<0.001
Random Effects			
σ^2	0.59		
T_{00} Participant	0.33		
T_{00} School	0.02		
T_{00} Task	0.01		
ICC Participant	0.35		
ICC School	0.02		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.049 / 0.406		
Proportion of active writing time - Interval 1			
Predictors	Estimates	CI	p
(Intercept)	0.14	-0.01 – 0.30	0.071
Grade 10	-0.06	-0.19 – 0.08	0.391
Grade 12	0.30	0.11 – 0.50	0.002
Gender Male	-0.22	-0.35 – -0.08	0.001
Genre INF	-0.18	-0.33 – -0.04	0.019

Random Effects	
σ^2	0.44
τ_{00} Participant	0.45
τ_{00} School	0.03
τ_{00} Task	0.04
ICC Participant	0.48
ICC School	0.03
ICC Task	0.04
Observations	2155
Marginal R^2 / Conditional R^2	0.037 / 0.560

Proportion of active writing time - Interval 2

Predictors	Estimates	CI	p
(Intercept)	0.01	-0.14 – 0.16	0.911
Grade 10	0.04	-0.10 – 0.17	0.593
Grade 12	0.26	0.07 – 0.45	0.007
Gender Male	-0.06	-0.19 – 0.07	0.382
Genre INF	-0.06	-0.20 – 0.08	0.402

Random Effects	
σ^2	0.49
τ_{00} Participant	0.42
τ_{00} School	0.02
τ_{00} Task	0.03
ICC Participant	0.44
ICC School	0.02
ICC Task	0.03
Observations	2155
Marginal R^2 / Conditional R^2	0.011 / 0.497

Proportion of active writing time - Interval 3

Predictors	Estimates	CI	p
(Intercept)	0.01	-0.10 – 0.13	0.830

Random Effects	
σ^2	0.44
τ_{00} Participant	0.47
τ_{00} School	0.04
τ_{00} Task	0.05
ICC Participant	0.48
ICC School	0.04
ICC Task	0.05
Observations	2155
Marginal R^2 / Conditional R^2	0.000 / 0.560

Proportion of pause time during production - Interval 1

Predictors	Estimates	CI	p
(Intercept)	0.00	-0.13 – 0.14	0.964
Grade 10	-0.11	-0.24 – 0.02	0.094
Grade 12	-0.09	-0.27 – 0.10	0.370
Gender Male	0.15	0.03 – 0.28	0.019
Genre INF	-0.02	-0.12 – 0.08	0.639

Random Effects			
σ^2	0.51		
τ_{00} Participant	0.39		
τ_{00} School	0.03		
τ_{00} Task	0.01		
ICC Participant	0.41		
ICC School	0.03		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.009 / 0.461		
Proportion of pause time during production - Interval 2			
Predictors	Estimates	CI	p
(Intercept)	0.13	-0.00 – 0.25	0.052
Grade 10	-0.14	-0.27 – -0.01	0.042
Grade 12	-0.02	-0.21 – 0.17	0.828
Gender Male	0.06	-0.07 – 0.19	0.365
Genre INF	-0.21	-0.29 – -0.13	<0.001
Random Effects			
σ^2	0.45		
τ_{00} Participant	0.44		
τ_{00} School	0.02		
τ_{00} Task	0.01		
ICC Participant	0.48		
ICC School	0.03		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.018 / 0.519		
Proportion of pause time during production - Interval 3			
Predictors	Estimates	CI	p
(Intercept)	0.12	-0.01 – 0.25	0.067
Grade 10	-0.16	-0.29 – -0.02	0.022
Grade 12	0.08	-0.12 – 0.27	0.443
Gender Male	-0.04	-0.17 – 0.09	0.551
Genre INF	-0.14	-0.23 – -0.06	0.003
Random Effects			
σ^2	0.51		
τ_{00} Participant	0.43		
τ_{00} School	0.03		
τ_{00} Task	0.01		
ICC Participant	0.44		
ICC School	0.03		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.014 / 0.484		
Total number of keystrokes typed			
Predictors	Estimates	CI	p
(Intercept)	0.23	0.08 – 0.38	0.004
Grade 10	-0.31	-0.45 – -0.17	<0.001
Grade 12	0.46	0.26 – 0.66	<0.001
Gender Male	-0.31	-0.44 – -0.17	<0.001
Genre INF	-0.13	-0.25 – -0.00	0.055

Random Effects			
σ^2	0.34		
τ_{00} Participant	0.50		
τ_{00} School	0.04		
τ_{00} Task	0.03		
ICC Participant	0.55		
ICC School	0.04		
ICC Task	0.03		
Observations	2155		
Marginal R^2 / Conditional R^2	0.100 / 0.665		
Number of keystrokes per minute - Interval 1			
Predictors	Estimates	CI	p
(Intercept)	0.06	-0.11 – 0.24	0.481
Grade 10	-0.21	-0.35 – -0.07	0.004
Grade 12	0.24	0.03 – 0.44	0.027
Gender Male	-0.12	-0.26 – -0.02	0.093
Genre INF	0.01	-0.16 – 0.19	0.892
Random Effects			
σ^2	0.35		
τ_{00} Participant	0.53		
τ_{00} School	0.04		
τ_{00} Task	0.06		
ICC Participant	0.54		
ICC School	0.04		
ICC Task	0.06		
Observations	2155		
Marginal R^2 / Conditional R^2	0.028 / 0.654		
Number of keystrokes per minute - Interval 2			
Predictors	Estimates	CI	p
(Intercept)	0.09	-0.05 – 0.24	0.198
Grade 10	-0.19	-0.34 – -0.05	0.010
Grade 12	0.07	-0.12 – 0.27	0.455
Gender Male	0.13	-0.01 – 0.27	0.073
Genre INF	-0.14	-0.26 – -0.03	0.022
Random Effects			
σ^2	0.39		
τ_{00} Participant	0.56		
τ_{00} School	0.01		
τ_{00} Task	0.02		
ICC Participant	0.57		
ICC School	0.01		
ICC Task	0.02		
Observations	2155		
Marginal R^2 / Conditional R^2	0.021 / 0.609		

Number of keystrokes per minute - Interval 3			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	-0.07	-0.22 – 0.08	0.363
Grade 10	0.06	-0.09 – 0.21	0.431
Grade 12	0.22	-0.01 – 0.44	0.063
Gender Male	0.25	0.12 – 0.38	<0.001
Genre INF	-0.08	-0.18 – 0.03	0.158
Grade10:GenreINF	-0.17	-0.30 – -0.04	0.012
Grade12:GenreINF	-0.11	-0.28 – 0.06	0.214
Random Effects			
σ^2	0.48		
τ_{00} Participant	0.44		
τ_{00} School	0.06		
τ_{00} Task	0.01		
ICC Participant	0.45		
ICC School	0.06		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.026 / 0.528		
Number of pauses per minute - Interval 1			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	-0.01	-0.09 – 0.08	0.869
Random Effects			
σ^2	0.58		
τ_{00} Participant	0.37		
τ_{00} School	0.02		
τ_{00} Task	0.02		
ICC Participant	0.38		
ICC School	0.02		
ICC Task	0.02		
Observations	2155		
Marginal R^2 / Conditional R^2	0.000 / 0.421		
Number of pauses per minute - Interval 2			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.13	-0.01 – 0.27	0.075
Grade 10	-0.22	-0.38 – -0.05	0.012
Grade 12	0.08	-0.14 – 0.30	0.452
Gender Male	-0.05	-0.24 – 0.14	0.617
Genre INF	-0.14	-0.23 – -0.05	0.004
Grade10:GenderMale	0.28	-0.00 – 0.56	0.053
Grade12:GenderMale	-0.18	-0.55 – 0.19	0.345
Random Effects			
σ^2	0.51		
τ_{00} Participant	0.42		
τ_{00} School	0.01		
τ_{00} Task	0.01		
ICC Participant	0.44		
ICC School	0.01		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.017 / 0.478		

Number of pauses per minute - Interval 3			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.25	0.12 – 0.37	<0.001
Grade 10	-0.26	-0.41 – -0.11	0.001
Grade 12	-0.07	-0.27 – -0.14	0.517
Gender Male	-0.16	-0.29 – -0.04	0.011
Genre INF	-0.21	-0.32 – -0.11	<0.001
Grade10:GenreINF	0.01	-0.14 – 0.15	0.931
Grade12:GenreINF	0.26	0.07 – 0.44	0.007
Random Effects			
σ^2	0.58		
τ_{00} Participant	0.37		
τ_{00} School	0.02		
τ_{00} Task	0.00		
ICC Participant	0.38		
ICC School	0.02		
ICC Task	0.00		
Observations	2155		
Marginal R^2 / Conditional R^2	0.034 / 0.424		
Mean pause time - Interval 1			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.06	-0.04 – 0.17	0.242
Grade 10	-0.12	-0.24 – -0.01	0.038
Grade 12	-0.04	-0.20 – 0.12	0.637
Gender Male	0.09	-0.02 – 0.20	0.121
Genre INF	-0.09	-0.17 – -0.02	0.014
Random Effects			
σ^2	0.77		
τ_{00} Participant	0.20		
τ_{00} School	0.02		
τ_{00} Task	0.00		
ICC Participant	0.20		
ICC School	0.02		
ICC Task	0.00		
Observations	2155		
Marginal R^2 / Conditional R^2	0.008 / 0.223		
Mean pause time - Interval 2			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.09	-0.03 – 0.22	0.152
Grade 10	-0.07	-0.19 – 0.06	0.316
Grade 12	-0.04	-0.22 – 0.14	0.673
Gender Male	0.04	-0.08 – 0.17	0.489
Genre INF	-0.18	-0.27 – -0.09	0.001
Random Effects			
σ^2	0.63		
τ_{00} Participant	0.33		
τ_{00} School	0.02		
τ_{00} Task	0.01		
ICC Participant	0.33		
ICC School	0.02		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.010 / 0.371		

Mean pause time - Interval 3			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	-0.01	-0.08 – 0.07	0.877
Random Effects			
σ^2	0.63		
τ_{00} Participant	0.34		
τ_{00} School	0.02		
τ_{00} Task	0.00		
ICC Participant	0.34		
ICC School	0.02		
ICC Task	0.00		
Observations	2155		
Marginal R^2 / Conditional R^2	0.000 / 0.368		
Produced ratio			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	-0.22	-0.36 – -0.08	0.003
Grade 10	0.26	0.12 – 0.41	<0.001
Grade 12	-0.49	-0.70 – -0.27	<0.001
Gender Male	0.55	0.42 – 0.68	<0.001
Genre INF	0.11	0.00 – 0.21	0.045
Grade10:GenreINF	-0.16	-0.28 – -0.04	0.008
Grade12:GenreINF	0.04	-0.12 – 0.19	0.627
Random Effects			
σ^2	0.39		
τ_{00} Participant	0.42		
τ_{00} School	0.05		
τ_{00} Task	0.01		
ICC Participant	0.49		
ICC School	0.05		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.125 / 0.604		
Number of transitions per minute between the sources - Interval 1			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.13	0.02 – 0.23	0.017
Grade 10	-0.09	-0.19 – 0.02	0.095
Grade 12	-0.11	-0.25 – 0.02	0.111
Gender Male	0.05	-0.05 – 0.15	0.330
Genre INF	0.10	0.01 – 0.20	0.040
Random Effects			
σ^2	0.42		
τ_{00} Participant	0.22		
τ_{00} School	0.00		
τ_{00} Task	0.01		
ICC Participant	0.35		
ICC School	0.00		
ICC Task	0.01		
Observations	2155		
Marginal R^2 / Conditional R^2	0.009 / 0.364		

Number of transitions per minute between the sources - Interval 2			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.25	0.16 – 0.33	<0.001
Grade 10	-0.04	-0.12 – 0.05	0.392
Grade 12	-0.12	-0.23 – -0.01	0.027
Gender Male	0.08	-0.02 – 0.18	0.114
Genre INF	-0.03	-0.12 – 0.06	0.563
Grade10:GenderMale	-0.07	-0.21 – 0.07	0.349
Grade12:GenderMale	0.20	0.01 – 0.39	0.037
Random Effects			
σ^2	0.23		
τ_{00} Participant	0.08		
τ_{00} School	0.00		
τ_{00} Task	0.01		
ICC _{Participant}	0.28		
ICC _{School}	0.00		
ICC _{Task}	0.01		
Observations	2155		
Marginal R ² / Conditional R ²	0.014 / 0.304		
Number of transitions per minute between the sources - Interval 3			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.18	0.11 – 0.26	<0.001
Grade 10	0.05	-0.03 – 0.13	0.211
Grade 12	-0.06	-0.15 – 0.04	0.259
Gender Male	0.10	0.01 – 0.19	0.027
Genre INF	0.08	0.00 – 0.15	0.055
Grade10:GenderMale	-0.21	-0.34 – -0.08	0.001
Grade12:GenderMale	-0.01	-0.18 – 0.16	0.921
Random Effects			
σ^2	0.24		
τ_{00} Participant	0.05		
τ_{00} School	0.00		
τ_{00} Task	0.01		
ICC _{Participant}	0.19		
ICC _{School}	0.00		
ICC _{Task}	0.01		
Observations	2155		
Marginal R ² / Conditional R ²	0.014 / 0.209		

Number of transitions per minute between the synthesis text and the sources - Interval 1

<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.13	-0.01 – 0.28	0.075
Grade 10	-0.10	-0.19 – -0.00	0.041
Grade 12	0.17	0.02 – 0.31	0.023
Gender Male	-0.27	-0.36 – -0.19	<0.001
Genre INF	0.26	0.10 – 0.42	0.003

Random Effects

σ^2	0.23
τ_{00} Participant	0.19
τ_{00} School	0.03
τ_{00} Task	0.05
ICC Participant	0.47
ICC School	0.03
ICC Task	0.05
Observations	2155
Marginal R^2 / Conditional R^2	0.077 / 0.580

Number of transitions per minute between the synthesis text and the sources - Interval 2

<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	-0.00	-0.13 – 0.13	0.963
Grade 10	0.02	-0.08 – 0.13	0.647
Grade 12	-0.10	-0.24 – 0.04	0.158
Gender Male	0.06	-0.04 – 0.16	0.262
Genre INF	0.28	0.15 – 0.42	<0.001

Random Effects

σ^2	0.26
τ_{00} Participant	0.26
τ_{00} School	0.01
τ_{00} Task	0.04
ICC Participant	0.49
ICC School	0.01
ICC Task	0.04
Observations	2155
Marginal R^2 / Conditional R^2	0.039 / 0.556

Number of transitions per minute between the synthesis text and the sources - Interval 3

<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.03	-0.05 – 0.11	0.469
Grade 10	0.07	-0.01 – 0.15	0.082
Grade 12	-0.11	-0.22 – 0.01	0.063
Gender Male	0.08	0.01 – 0.16	0.036
Genre INF	0.29	0.22 – 0.35	<0.001

Random Effects

σ^2	0.18
τ_{00} Participant	0.15
τ_{00} School	0.01
τ_{00} Task	0.01
ICC Participant	0.45
ICC School	0.01
ICC Task	0.01
Observations	2155
Marginal R^2 / Conditional R^2	0.072 / 0.511