Assessing Rater Reliability and Assessment Accessibility of a Newly-Designed, High-Stakes, Documented Writing Assessment in The Netherlands

Faculty of Behavioural Management and Social Sciences

University of Twente

Victoire Nijland

Supervisor 1: Dr. E.C. Roelofs

Supervisor 2: Dr. J.W. Luyten

The Road Not Taken

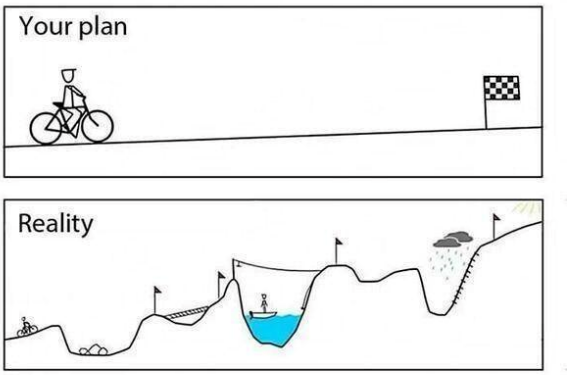
Two roads diverged in a yellow wood,
And sorry I could not travel both
And be one traveler, long I stood
And looked down one as far as I could
To where it bent in the undergrowth;

Then took the other, as just as fair,
And having perhaps the better claim,
Because it was grassy and wanted wear;
Though as for that the passing there
Had worn them really about the same,

And both that morning equally lay
In leaves no step had trodden black.
Oh, I kept the first for another day!
Yet knowing how way leads on to way,
I doubted if I should ever come back.

I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference

Robert Frost, 1916



Available from: https://www.researchgate.net/figure/The-plan-reality-dichotomy_fig4_358416173 [accessed 12 Jun, 2024]

I had a clear destination from the beginning, educational measurement in assessment. So, I took to the 'EST road', took a few wrong turns; got lost a couple of times; collided with fellow travellers; and found some true companion travellers!

For me, taking this course was not about getting a diploma, but about learning and becoming knowledgeable in educational measurement and the confidence that I, as a linguist who had never taken a maths course after secondary school, could do this. Beside the fact that I did it, I also got reassured that my beliefs about teaching and designing educational materials were not weird, they were just ahead.

So, thank you for guiding me along the way and patiently waiting at the destination, Erik. You were a patient and kind tutor. Thank you, Kirsten and Ester for your seemingly, unshakable belief in me during the final part of the journey, for keeping me from wandering off too far. Thank you, Lonneke and Irma for your conversations about writing and participating in this study.

Thank you, Renzo and Hjalmar for just being there. We're done for now, I can go back to doing the mum and wife stuff again – no more uni-studying for another 20 years, I'll have to save some money first. 😉

1 Content

2 Abstract

This study explored how a newly-designed, norm-referenced, documented writing assessment was perceived by higher secondary education students in their final year and (teacher)raters before, during and after administration.

Six raters assigning ratings to nine sub domains of one writing task based on 26 performances presented a G-coefficient of .81 for agreement without training or exemplars. A G-coefficient of .68 was found for two teacher raters. Further research should determine if training or instructions improve agreement levels. Qualitative results showed that raters rely on a) their beliefs about the concept and its domains, and b) rating experience in the form of scoring stages to assign ratings that discriminate between students' writing abilities. Thus, raters could identify differences in performances, and they are comfortable applying the criteria. Despite varying rater beliefs and methods, six raters rated consistently, meaning that, this sample of raters have a mutual understanding of the concept documented writing and its domains.

This study also found that assessment accessibility factors, 1) effectiveness of instructed writing strategies, 2) difficulty level of sources in documentation file and 3) assessment administration setting significantly correlated on obtained writing scores. Hence, when making decisions on students' relative standing in writing ability using this writing task, the impact of these factors for certain groups of students should be considered when determining pass/ fail grades. Also, the design of a standardized writing assessment along with its support materials should be aligned with the construct based on language abstraction, language level and genre.

Keywords: educational measurement; high-stakes documented writing assessment; rater reliability; rater agreement; assessment accessibility

3 Introduction

Being able to write is essential for higher education and work life (Deane, 2011; Graham et al., 2013). That is why writing skills for Dutch language for pre-university education were assessed in a second exam session in the Dutch centralised exams. However, due to low reliability in ratings and dissatisfaction about what was assessed in the centralised writing exam session, it was removed from the centralised exam in 1998 (Schoonen, 1997; Rooijackers, 2007).

Currently, Dutch pre-university students manifest their writing ability and their ability to revise scripts based on received feedback through school-based, documented assessments in either expository or argumentative writing. These school-based assessments are designed and scored by individual school departments (Ekens & Meestringa, 2013). For which assessment matrices, rating models, and professional conferences about marking are often lacking, resulting in inadequate ratings of school-based writing exams and variety in writing ability among students entering higher education (Bouwer et al., 2022; Ekens & Meestringa, 2013; Nederlands Nu! & Sectie Bestuur Nederlands Levende Talen, 2018; van der Leeuw & Meestringa en Ravesloot, 2012).

To remedy current variety in students' writing abilities entering tertiary education, a call for revision of the national writing syllabus emerges among teachers, syllabi designers, and Dutch language experts; The (re)integration of writing in the centralised assessment in the final year of secondary education (Hendrix & van der Westen, 2018; Nederlands Nu! & Sectie Bestuur Nederlands Levende Talen, 2018; Rooijackers, 2007), thus reintroducing writing assessment as a norm-referenced assessment.

However, assessing writing performance in a high-stakes, centralised setting causes a paradox. For one, because innovations in language education, and particularly the view of language proficiency as part of an integrated skill set (Chan & Yamashita, 2022; Deane, 2011; Nederlands Nu! & Sectie Bestuur Nederlands Levende Talen, 2018) may not fit the strict separation of language skills in current Dutch central examinations. Moreover, the

introduction of new assessment methods tends to result in lower score reliability and an increase in sources of measurement error (Gebril, 2010; Lee & Kantor, 2005). Therefore, considerations about warrants for valid conclusions about high-stakes writing assessment need to be made, such as the number of assessments per student, the number of raters per performance and the method in which the assessment should be scored (Hendrix & van der Westen, 2018). In short, in envisioning an assessment procedure for a centralised writing exam that reflects writing as complex integrated language skill set, and that is expected to warrant national comparability of exam results, assessment validation becomes even more important.

Validity in writing assessment is 'the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests (American Educational Research Association. et al., 2014). Writing is a complex skill and its assessment manifests itself in performance tasks that are rater-mediated. Assessment consistency, or reliability, is measured by the sources of error in assessment scores and their causes, such as the writer, the task itself, and raters, as well as the scoring procedures (Hamp-Lyons, 2012). This means that reliability in writing assessment is not based on the writer alone and final scores will always be distorted by task and rater. Therefore, this study explored the reliability and accessibility of a newly-designed, high-stakes, documented witing assessment task by obtaining G-coefficients for agreement and reliability, supplemented by qualitative rater and student data to provide insights into the rater perceptions of the rating process as well as student perceptions about assessment accessibility before and during assessment administration.

4 Theoretical Framework

4.1 Generalizability of Writing Assessment Scores

Reliability of a high-stakes writing assessment shows the degree to which scores for writing ability can be generalized across raters and predefined task dimensions, which are considered to span the writing construct, including genre and purpose. However, with a performance assessment of a complex skill such as writing ability, score generalizability is

threatened by the factor raters, possibly due to the different use of rating procedures, and by task-specific variation itself , whereas the main valid source of score variation should be the writer, manifesting their true writing capability during the writing task performance.

First, large-scale high-stakes writing assessments are human rater-mediated performance tasks prone to bias, as human rating means interpretating the rating criteria, thus some form of reliability reduction due to raters must be considered (Bouwer et al., 2022; Elosua, 2022; Wind & Engelhard, 2013). Rater reliability refers to the first type of inference in the argument-based approach to validity for performance assessments as introduced by Kane (2013). The three inferences in argument-based validity are, a) the scoring inference, in which scores are obtained from performances, b) the generalization inference, in which scores are interpreted as a reflection of the test domain, and c) the extrapolation inference which extends the interpretation into real-world performance domains (Kane, 2013; Kane et al., 1999).

Warrants for making valid inferences about obtaining observed scores from rating are a) the criteria used to score are appropriate and have been applied as intended, and b) assessment administration conditions are comparable with its purposes. The inference for the second step, the generalization inference, means that based on the tasks scores administered a statistical generalization can be made towards a defined universe of tasks. In short, the tasks applied should represent the predefined domain of all possible tasks. The third inference, the extrapolation inference, consists of the extrapolation of the target score towards the target domain. Evidence is needed to demonstrate that (writing) performance on the assessment tasks is predictive or strongly related to writing in practice. In short, to back-up the credibility of the inferences, both theoretical and empirical evidence is needed (Kane, 2013; Kane et al., 1999).

The argument-based approach to validity employs reliability as one of its assumptions for the generalizability inference. Depending on the assessment, reliability relates to whether an assessment is consistent across different versions, across different groups, or across different raters (Greenberg, 1992; Hamp-Lyons, 2012; Livingston, 2018). In writing

assessment, reliability is measured by consistency across raters, using measures of interrater agreement and rater reliability. More specifically, the measures refer to the extent to which independent raters consistently assign scores to domains according to the rating procedures.

An issue that touches upon extrapolation of writing assessments is what can be summarized as ‘construct reduction’. Critics state that writing assessments often lead to construct reduction, for in a high-stakes setting, it is hardly possible to assess all aspects and domains of writing. Thus, the tendency in writing assessment to narrow the concept ‘writing ability’ to one task (Beck & Jeffery, 2007; Deane, 2011; Deane et al., 2008; Moss, 1994; Slomp, 2012; Slomp & Fuite, 2004; Van Den Bergh & Meuffels, 2000; Wiggins, 1994), threatens inferences for both generalization to the assessment domain and extrapolation to the target domain (Kane et al., 1999; Lederman, 2018). However, although these ‘one-task’ assessments render lower reliability or generalizability, they may provide better extrapolations about the student’s real world writing abilities (Lederman, 2018).

Another issue that relates to accessibility of high-stakes writing is that certain test-takers may have little content knowledge of the subject required in the writing task which may result in construct reduction and, thus cause bias in assessment results. Additionally, when writers do not have access to content information and little time for processing due to administration setting, this results in test-takers focusing on the syntactic and semantic features of writing instead of content (Deane, 2011; Flower & Hayes, 1981).

4.2 Writing Models and Frameworks

Writing models have been presented in frameworks that map the theoretical aspects of writing for research and the practical aspects of writing, based on the theoretical models, for teaching and assessing. In both writing is considered a complex task that requires solving difficult problems during various cognitive and metacognitive processes (Flower & Hayes, 1981) in a social context. Since the 1980s, conceptual and cognitive models for writing have been devised and revised (Cahill et al., 2013; Hayes & Flowers, 1983).

A cognitive writing model was presented by Flowers and Hayes (1981). It maps a complex interaction of cognitive load on working memory during writing. In this model, the composing process consists of three main processes: planning, translating, and reviewing. The writer monitors the composing process internally. Goal setting pertains to the process in terms of what the writer intends to do with the text in the next few steps. Throughout the writing process, the reviewing process takes place; a writer rereads, sets new goals, and alters previously written text. The task itself is the writer's external goal and triggers long-term memory for knowledge of topic (content knowledge), and the execution of automated composing processes (Huot 1990; Moss, 1994; Bazerman, 2015). The act of writing, thus, places a high demand on cognitive load, which was postulated in the Capacity Theory for Writing (McCutchen, 1996). Demands on long-term memory serve for retrieval of content knowledge, knowledge about structure, lay out, and linguistic features of the text and at the same time, working memory executes the processes of planning, translating, and reviewing.

From a practical perspective in teaching and assessing, writing ability has been measured in terms of product, process, and genre-context perspective (Rijlaarsdam et al., 2005). In product-approached writing, focus is on the written product of a writing task; in process approached writing, focus is on the writing processes evoked by the task; and the genre-context based approach combines product and process but places them in the context of social discourse (Slomp, 2012). Bazerman (2015) defines writing ability as "a complex social participatory performance in which the writer asserts meaning, goals, actions, affiliations, and identities within a constantly changing, contingently organized social world, relying on shared texts and knowledge [p18]." In short, writing ability is a complex skill that not only demands knowledge of linguistic skills, and metacognitive processes, but also a social context for output and cognitive and conceptual input. These three elements, processes, knowledge, and strategies apply to any genre of writing, but each genre requires different specifications as to what processes, knowledge and strategies are invoked (Ekens & Meestringa, 2013; Schuurs, 2021).

4.2.1 Conceptual Writing Assessment Frameworks

To make valid generalizations from the task domain, the writing task reflecting the test domain should assess the writing skillset appropriate for the performers being assessed. However, although conceptual and theoretical models of writing exist, writing ability, and especially integrated writing, is still a topic of ongoing interest in research.

Deane et al. (2008) suggest that the skills needed for narrative, expository and argumentative writing are based on knowledge, verbal reasoning, social evaluative, and linguistic and rhetorical skills and their inherent processes and strategies. This view is further integrated in Deane's Cognitive Model for Writing Proficiency (2011), in which reading, writing and critical thinking are different activity types of an underlying skill-set for literacy (Deane, 2011). In this view, literacy consists of five skill layers; social, conceptual, textual, verbal, and lexical / orthographic, which on their own require different modes of thought and representation for interpretation, deliberation, and expression in the literacy process. In short, 'skilled writers combine efficient receptive and expressive skills with appropriate and effective reflective strategies" (Deane, 2011, p.11).

The amount of working memory a writer can free up for the various writing processes, skills, and strategies required for writing depends on which process requires the highest cognitive load; whereas a novice writer, for example, may not possess the cognitive ability to focus on both transcription and knowledge retrieval at the same time, an experienced writer will have automated orthographic processes and cognitive resources for reviewing processes; in other words the transformation from knowledge-telling to knowledge-transformation (Becker, 2006; Deane et al., 2008; Galbraith & Rijlaarsdam, 1999; Mccutchen, 1996; Scardamalia & Bereiter, 1987). Thus, narrative writing seems to demand less cognitive load than expository or argumentative writing since the writer merely draws upon their own knowledge. In this light, narrative writing, expository writing, and argumentative writing may reflect various stages of writing proficiency, for each requires different processes, strategies, and skills (Deane et al., 2008; Chan & Yamashita, 2022).

4.2.2 Current Writing Assessment Frameworks

Academic skill sets are assessed through knowledge transforming, argumentative writing (Elander et al., 2006), therefore, pre-university students in their final year of secondary education should be proficient in at least the basics of argumentative writing and qualifying curricula should reflect these concepts or dimensions of writing. A short overview of writing assessment frameworks from European and Anglo-Saxon countries gives an idea of what genre of writing, what type of task and on what criteria these tasks are assessed among pre-university students aged between 17 and 19 (Curriculum for Norwegian, 2020; Elf & Troelsen, 2021; Perelman, 2018; Skar & Aasen, 2021).

Table 4.1

Pre- University L1 Writing Frameworks in European and Anglo-Saxon countries

	Integrated or Independent Writing	Student Age in Years at Administration	Genre of Writing Assessed	Administration Time	Assessed on
Common Core Standards, USA	IND	17-18	expository, narrative, argumentative writing	2 x 30 mins	language control effective for purpose and audience
National Curriculum – GAT / ACSF, Australia	INT	15-16	expository, argumentative, and narrative writing		knowledge of text, grammar, word, and visuals
United Kingdom – A' Levels English Language	INT	17-18	academic essay writing	integrated skill in two sessions	audience, purpose, genre, and mode, exploration of language in its social and geographical contexts
The Norwegian Framework	INT	19		120	use of language argue on interdisciplinary and subject-related topics, coherence, punctuation, spelling, different genres
The Danish Framework	INT	16-17	argumentative, expository writing		

Table 4.1 shows that, at pre-university level, the genre of writing assessed is either all three types of writing; narrative, expository, and argumentative writing; or expository and argumentative writing. These tasks are assessed on at least the following three domains:

effective use of language for audience and genre or purpose, accurate use of language, and cohesion.

The genres are, apart from the Common Core Standards, assessed by integrated writing tasks. In integrated writing tasks the writer processes (provided) sources to select, organize and integrate information into their writing task (Chan & Yamashita, 2022; Knoch & Sitajalabhorn, 2013).

4.2.3 Dutch Framework for Writing

The Dutch national syllabus (College voor Toetsen en Examens, 2021) sets performance criteria for centralised and school-based summative assessment in the final year of secondary education, Referentie Kader taal en rekenen [Framework language and maths] (Meijerink et al., 2009) has been implemented and used since 2010, and serves as a benchmark to indicate key levels with descriptors in the development of literacy and mathematics throughout primary and secondary education. For the final years of secondary education, especially levels 3F and 4F, which roughly compare to CEFR B2 and C1 level, preparing students for academic education, descriptors are alike for writing. Table 4.2 presents an overview of the descriptors for all levels.

Table 4.2

Dutch Framework Descriptors for Writing all Levels

	1F	*2F*	*3F*	*4F*
Common descriptor	Can write short, simple texts about daily subjects or about subjects from everyday life.	Can write coherent texts with simple, linear structure, about a range of familiar subjects in work life or from social science nature.	Can write detailed texts about subjects from work life or social science, in which information and arguments from several sources are integrated and evaluated.	Can write well-structured texts about a wide range of subjects from work-life or training. Can emphasize issues, elaborate arguments, and support them with reasons and examples.

This framework aligns with Deane et al.'s conceptual identification of the three genres of writing, as well as what is assessed in writing internationally: narrative, expository and argumentative writing (2008). Like Deane et al.'s (2008) views, proficiency levels build from

narrative writing at the lower level to argumentative writing at the higher level. The framework presents performance objectives for level 1F to 4F per genre. Additionally, it presents criteria per level for cohesion, effectiveness in purpose and audience, use of language, vocabulary, and readability.

The national syllabus commissioned by the College van Toetsen & Examens [Committee of Assessment & Exams] provides similar descriptors for writing. In table 4.3 the descriptors for writing at 3F level, havo, can be seen.

Table 4.3

National Syllabus Descriptors for Domain C, Writing Havo.

Domain C Writing	*The candidate can, for the purpose of writing a documented, expository, and argumentative text,*
	- select and process relevant information. - present the information effectively taking in account purpose, audience, and type of text, as well as conventions for written language. - revise concepts of written text based on received feedback.

The criteria reflect the construct of integrated writing (Plakans & Gerbril, 2013) in that it supposes knowledge about conventions and purpose of text genre, while at the same time evoking writing processes such as the selection, organization, integration, and revision of ideas to produce a written argumentative or expository composition. A documented essay task is a type of integrated writing, in which students are provided with, or search for written sources to mine, select and organize ideas and translate these into a coherent composition, keeping in mind the conventions of the genre and the social elements of writing.

4.2.4 The Construct Integrated Writing

Integrated writing incorporates the dimensions of the conceptual framework for writing, which suggests that writing is a complex skill drawing on several skills, processes, and strategies in a social context. To perform integrated writing, the writer receives either audiovisual or textual input from which information is selected and transformed into a written output. Yet, integrated writing is considered a ‘construct’ separate from independent writing in which the writer draws on memory to perform the task (Chan & Yamashita, 2022; Ohta et al., 2018). Although some domains and processes of integrated writing are like independent

writing, such as organisation, cohesion, (Chan & Yamashita, 2022; Plakans & Gebril, 2017) and syntactic complexity (Chan & Yamashita, 2022), other domains seem to be exclusively for integrated writing. These particular domains may cause variability in students' integrated writing scores. First, Chan & Yamashita (2022) found that source integration, in terms of organization and paraphrasing, was a sub domain correlated to integrated writing. Also, topical knowledge is an important predictor for the quality of integrated writing products (Deane, 2011). By providing sources for the integrated writing task or standardizing topical knowledge input, especially writers having small breadth and depth of vocabulary benefit and, thus, decrease variability in writing scores (Schoonen, 2005; Chan & Yamashita, 2022). This could be remedied by providing sources for integrated writing tasks that are selected based on similar topic specificity and rubrics that include rating the use of sources (Homayounzadeh et al., 2019).

A second cause for variability in integrated writing tasks are its assessment procedures. When integrated writing is assessed in high-stakes, standardized, timed sessions, it tends to lack task authenticity. It does not reflect the process of reviewing and revising in real-life situations. However, Kim et al. (2018) find that feedback does not have a significant effect on integrated writing scores. What does, is the use of digital devices, as these provide the student with opportunities to easily reorder and revise concepts from previous drafts, even in one sitting (Kim et al., 2018). It follows from this, that in assessment, test-takers do not necessarily benefit from external reviewing, but from the social skill of the writer taking on the role of critical reader (Deane et al., 2008) and being allowed to work with devices.

A third process for integrated writing that may cause variability in writing scores when administered in a timed high-stakes administration is planning. McCutchen (1996) shows that planning writing frees up cognitive load in working memory and results in better final writing performances. This aligns with the idea that the more knowledgeable a writer is on the topic of the writing task, the less cognitive load is lost on processes such as retrieving, selecting, and organizing ideas (Becker, 2006). Therefore, in this study organization of ideas is

controlled by having students create a skeletal outline, demanding a form of source integration before the assessment task, which frees up cognitive load to focus on the writing aspects of the task.

4.3 Unwanted Variability in Integrated Writing Scores

Construct-irrelevant variance in writing scores is the variability in scores that is caused by factors that are not related to the construct of integrated writing. This construct-irrelevant variance could be caused by raters' interpretation of the rating criteria, by their methods of rating, their backgrounds, experiences, and beliefs. Another possibility of construct-irrelevant variance stems from assessment accessibility. If students experience barriers in the preparation for and during the administration of the assessment, this could impact students' ability to manifest their true writing abilities (Elliott et al., 2018).

4.3.1 Raters

In contrast to practice in other countries, Dutch teachers function as raters of their own students in Dutch centralised exams. After a teacher rates their own students, their scores are then vetted by a teacher from another school. After conference, these two teachers finalise student scores. If writing is reintroduced and moved from the school-based exam setting into the current central exam setting, secondary schoolteachers are trusted to rate their own students' high-stakes writing assessment.

Most rater research is collected from trained and qualified raters, rating in large validated high-stakes settings. Research concerns issues such as a rater's interpretation of complex rating scales, their interpretations of the wording of a scale, and indistinction between scales (Heidari et al., 2022; Ono et al., 2019). Even rater experience and rater background can result in variability in ratings among raters who rate anonymous performances (Bouwer & Koster, 2016; Deygers & Van Gorp, 2015; Meadows & Billington, 2010; Palermo, 2022).

Research about teachers functioning as raters, who are not specifically trained, qualified, monitored and rate in groups is scarce. The studies that do investigate teachers as raters of writing are in a primary or lower secondary school setting. Gamaroff (2000) studied

teachers holistically rating primary school language tasks and found significant variability between teacher raters. Variability was due to rater interpretation and weighting of criteria in rating procedures. Jönsson et al. (2021) studied 42 teacher raters holistically and analytically rating four anonymized tasks from four 12-year-old students spread across the semester and present an overall grade with justifications. Rater agreement was between 52 and 67%, which might be explained by teacher raters' personal weighting of certain criteria in the rating rubrics. Like (Jönsson et al., 2021), Brookhart (2013) concluded that teacher raters may affect scores due to personal beliefs about writing and their interpretation of assessment criteria, which is like Graham's conclusion that score variability from teacher raters could be due to teacher beliefs about writing and teacher beliefs about students' skills and motivation (2019). However, even trained, and qualified raters affect scores by personal beliefs and preferences about writing and rating criteria (Eckes, 2008).

Skar & Jølle (2017) investigated rater reliability of eight especially trained teacher raters for 2 high-stakes writing tasks performed by 25 students. Raters rated a grade-nine narrative and expository text, assigning ratings analytically on to six domains each consisting of six-point scale. All 8 raters rated the 50 performances. An intraclass correlation coefficient (ICC) of .93 based on all ratings of the two tasks was obtained. With training, teacher raters could consistently rate performances if they used the same rating criteria. Brown et al. (2004) find similar reliability levels for classroom teachers rating lower secondary school standardized writing tasks. In this study, even with little training, teachers reach high adjacent agreement of between 70-80% and acceptable reliability rates between .70 and .80 using analytic rating criteria.

Therefore, relying on research that two human raters considerably improve the reliability of rating high-stakes performance assessments (Bouwer & Koster, 2016; Gebril, 2009; Johnson et al., 2005; Livingston, 2018; Wind, 2019) and that background in raters does not affect variability in scores, teachers function as raters in this study.

4.3.2 Rating Criteria

Alignment between domains of the construct and rating criteria prevents construct-irrelevant variance in integrated writing assessment scores. Thus, a rating rubric should be designed to enable raters to assign scores to different, but mutually exclusive domains, such as grammar, cohesion, and structure. Analytic multi-trait scale rubrics are considered most reliable in writing assessments (Chan et al., 2015; Jönsson et al., 2021; Ohta et al., 2018; Schoonen, 1997; Schuurs, 2021). However, analytical scale rubrics are open to rater interpretation due to potentially longer scales resulting in an increase in centrality, scores centred around a certain score, and misfit, consistently rating a performed test higher or lower (Heidari et al., 2022; Malone, 2013). Yet, raters, when given a choice about rating procedures, were found to be more positive towards analytic scales because of its usability in conference after rating (Schoonen, 1997; Zou, 2022) and for feedback purposes (Bouwer et al., 2023). Also, descriptive, and distinguishable rubrics should be designed for the various genres of writing (Bouwer & Koster, 2016; Ekens & Meestringa, 2013; Knoch & Sitajalabhorn, 2013) to prevent rater-effects (Humphry & Heldsinger, 2014).

4.3.3 Assessment Accessibility

Another cause of variability in high-stakes writing scores is assessment accessibility. If assessments are not accessible, they will prevent test-takers from manifesting their true abilities. Elliott et al. (2018) define assessment accessibility as a process that

> 'involves removing obstacles that limit students' opportunities to learn the intended and tested curriculum, deny or disrupt their receipt of individualized accommodations for learning and testing, and reduce the degree to which tests provide accurate information about their knowledge and skills (p.1).

According to Elliot et al (2018) assessment accessibility for students involves the a) opportunity to learn such as instructional time, content coverage, and quality of instruction, b) academic enablers which includes skills, attitudes, and engagement behaviours, and c) assessment administration accessibility. If assessment accessibility is high, variance due to assessment accessibility is low, meaning that variance due to performance variability is high.

Research Questions

This study explores the reliability and assessment accessibility of a documented writing task, the specific research questions are summarised in table 4.4. This study explores if students' writing performances are scored consistently across multiple independent raters. In addition, raters' experiences regarding their preparation for the rating process and their perceived procedures in arriving at scores are explored through interviews. Finally, this study explores the assessment accessibility of the newly designed writing assessment, by focusing on students' perceived opportunity to learn, their attitudes and engagement behaviours, and their evaluation of the assessment administration accessibility. This is done through self-perception questionnaires. Relationships between assessment accessibility on the one hand and writing scores on the other hand are explored.

Relevance

The results of this study aim to contribute to insights on rater agreement of a newly-designed, documented writing assessment in a high-stakes setting. By employing a G-coefficient, a comparable measure of rater agreement is introduced. By focusing on rater backgrounds, preparations and procedures, this study will also contribute to research on rater background.

Also, the context of this study differs from previous ones, as its context shifts from primary and lower secondary (Bouwer et al., 2015; Schipolowski & Böhme, 2016)into a high-stakes and upper secondary school setting. Thus, this study presents an idea of student perceptions of assessment accessibility in terms of opportunities to learn, attitudes, and administration accessibility.

Finally, the practical aim of this study is to contribute to the development of guidelines for the design, administration, and rating procedures of an integrated writing assessment in the Dutch centralised examination context.

Table 4.4

Overview of Research Questions

	Quantitative Research	Qualitative Research
Rater Reliability	• To what extent do raters assign corresponding scores when rating a documented writing task performed by students in their final year of pre-university education in a high-stakes assessment setting, using a rating procedure designed for this purpose?	• Which aspects of the assessment procedure for the documented writing assessment support or hinder raters in arriving at valid interpretations of writing products and deriving a score based on these interpretations?
	• To what extent do teacher raters' scores of a high-stakes writing assessment show rater agreement? • To what extent do external raters' scores of a high-stakes writing assessment show rater agreement?	• How do raters prepare themselves to rate documented writing assessments? • What hindrances and supports do raters experience in assigning scores for documented writing assessments?
Student Perceptions	• How do students perceive the documented writing assessment procedure in terms of assessment accessibility?	
	• To what extent were students prepared for the new writing assessment task (perceived opportunities to learn)? • To what extent did students' attitudes towards writing and their engagement for writing present obstacles in manifesting their true writing ability for this designed writing task? • To what extent did students perceive the assessment administration as accessible?	

5 Research Design

This study explored to what extent a newly-designed, high-stakes, documented, writing assessment is scored consistently by raters. In addition, it investigated to what extent score variability can be explained by differences in accessibility of the writing assessment task. It investigated how the scoring inference holds for a sample of final year, pre-university students and a group of 6 raters.

Variance components analyses using raw data from ratings in a crossed design were employed to obtain G-coefficients for both absolute and relative agreement (Brennan, 2010; Heuvelmans & Sanders, 1993). Qualitative data collected from raters through an interview and questionnaires were analysed to explore how raters arrived at their scores, from which potential explanations can be yielded for observed score consistency or inconsistency. Additionally, quantitative data collected from students through self-perception questionnaires were used to explore to what extent assessment accessibility affected variability in scores. Assessment accessibility was investigated through students perceived opportunity to learn, their attitudes toward writing and their engagement behaviours, and their perceived assessment accessibility during administration.

5.1 Participants

5.1.1 School, teachers, and students

The participating school was selected through a convenience sample found after a call for participation was published in an online *Facebook* Group for teachers of Dutch. One exam year of a secondary school participated in this study, enrolling a total of 68 students. The students were in their final year of Havo (higher general education). Students in the sample originated from three classes and two teachers. Teachers that participated with their classes also functioned as raters for the writing assessment tasks and were interviewed after the rating procedures.

A total of 6 raters rated the writing tasks, including the two teachers. Four external raters were selected through a convenience and snowball sample of teachers from the personal network of the researcher and through the networks of assessment experts at the

Dutch Institute for Educational Measurement (Cito). Raters were recruited and paid €10 per rated performance. All raters were female and had at least 5 years of experience in teaching Dutch to exam years at secondary schools. This also means that the sample had experience with rating centralised high-stakes exams. Table 5.1 presents the background characteristics of the raters for the assessment. This table includes the two teacher raters.

Table 5.1

Demographic Features of Rater Participants

Rater	Teaching experience in years	Hours spent rating	Role	Education Levels	Focus writing education
Rater 1	25	n.a.	teacher	havo 3, 4, 5	n.a.
Rater 2	5	n.a.	teacher	havo 3, 4, 5	sentence structure
Rater 3	8	5	external	All levels havo and vwo	spelling, use of language, cohesion, references, originality
Rater 4	41	13	external	3, 4, 5 and 6 vwo	spelling, creative use of language, cohesion
Rater 5	41	7	external	predominantly 3, 4, 5 havo	spelling, use of language, references, genre conventions
Rater 6	11	7	external	4 mavo, 4, 5 havo, 5 vwo	use of language, cohesion
Mean	21.8	8			
SD	9.9	1.4			

5.1.2 Designed Materials for the Documented Writing Assessment

Materials were designed to restrict variability in writing scores. These materials included 1) lesson series to enable students to learn about writing strategies and to support teacher in writing instruction, 2) a documentation file, to provide students with similar input for content knowledge adjusted to their educational level, 3) an assessment task, and 4) a rating rubric.

Lesson Series. A lesson series of 9 lessons was designed to provide teachers with tools to instruct effective writing strategies. It contained 9 strategies to teach writing (Graham & Perin, 2007), such as information about genre conventions, giving and receiving peer feedback on drafts, creating a skeletal outline, modelling of problem-solving practices in writing, collaborative writing, and summarization. Each lesson was designed to last 45 to 50 minutes, defined by goal, and provided with instruction method. The lesson series was designed to instruct about how to write additional genres, such as letters and reports, with

the intention of familiarising students with the topic of the assessment. However, teachers were free to adapt the lesson series to meet their own and students' needs, therefore, not all elements were applied during instruction and some elements, such as citing sources, were added to the lesson series by teachers themselves.

Documentation File. Differences in students' content knowledge for the topic-to-be written about, causes variance in student scores. Therefore, a documentation file was compiled to provide students with content knowledge on the to-write-about topic (Schoonen, 2005) The file contained five expository and argumentative sources on the subject gender-neutral language to be handed out among students at the start of the semester. Accessibility of the sources was checked with the online tool Accessibility and Editor Tool in MS Word. The results of these accessibility checks of the sources can be found in table 5.2. The results indicated the sources were at CEFR B2/ C1 level, which translates to 3F and 4F level. However, the tool Accessibility categorizes texts based on semantic level instead of syntactic complexity (Kraf et al., 2011). Therefore, two language experts, one assessment expert and two qualified and experienced L1 teachers were asked to assess the accessibility of the sources in the documentation file keeping bearing in mind the sample of students. All agreed the sources were at a suitable level for the sample group of students. On top of that, exercises in the lesson series were designed to familiarize students with the contents of the documentation file.

Table 5.2

Accessibility Features of the Documentation File Students Used to Prepare for the Writing Assessment Task

source	estimated reading level (CEFR)	word count	M word length	M word count per sentence	genre of text
Source 1	B2*	423	5.7	21	expository
Source 2	B2	270	5	22	argumentative
Source 3	B2/ C1**	590	5.5	17.7	argumentative
Source 4	B2	815	5.1	17.4	expository
Source 5	B2	383	5	14.9	expository

Note * can read articles and reports about contemporary issues, in which the writers adapt a specific perspective or attitude. I can understand contemporary literary prose.

** can understand long and complex factual and literary texts and appreciate the different styles. I can understand specialized articles and long technical instructions, even when they do not concern my field of knowledge/ interest.

The Assessment Task. An open-ended expository writing task was designed for the purpose of this study, based on the Dutch Framework of Reference for Language (Meijerink et al., 2009)and the descriptors in the CvtE syllabus for domain C, writing (College voor Toetsen en Examens, 2021b). The instruction of the task was as follows:

Context

The National Action Committee Students is an organisation from, for and by students. Because of NACS conversations are not only about students, but also with students. NACS organises different activities, informs and represents students. NACS has an opinion about everything that has to do with secondary education, thus also about gender neutral language.
You have read the recommendations from NACS about gender neutral language. Write an expository essay about gender neutral language at school. The essay will be published on the publicly accessible website of NACS.

The Task

- *You will write an expository essay of (at least) 500 words;*
- *You choose a perspective and text structure matching this subject;*
- *You provide your essay with an appropriate title;*
- *You use at least two sources.*

For the task, the following applies:

- *Apply text genre and audience as given in the task.*
- *Make sure to create clear paragraphs.*
- *Mind sentence structure, spelling, and rhetorical devices.*

In preparation of the assessment task, students created a skeletal outline for their expository writing task based on the sources in the documentation file and at least two additional sources they looked for individually. The skeletal outline consisted of no more than 200 words including references. Students were allowed to bring the skeletal outline to the assessment administration to use as support.

5.2 Instruments

5.2.1 Rating Rubric

The rating rubric was based on Dutch writing rating rubric created by the Stichting Leerplan Ontwikkeling (SLO), the Dutch national expertise centre of syllabus development (Ekens & Meestringa, 2013) which was originally designed in cooperation with teachers of Dutch. The SLO rubric can be found in Appendix D. For this study, the SLO rubric was adapted in deliberation with assessment experts from Cito and the teachers taking part in the study. The new version of the rubric can be seen in table 5.3. The original number of four domains was reduced to three. Student instruction in the assessment task informed students about text genre and audience making the original domain C, text is 'appropriate for genre and audience' redundant. Moreover, because certain sub domains in the original domain 'appropriate for genre and audience' tended to overlap with sub domains in for example domain A, Coherence, and domain D, Presentation, they were removed from the rating form as to prevent rating aspects double. Additionally, because the expository writing task demands a certain structure, the SLO rubric domains A, Coherence, and B, Subject, were reviewed and recategorized. The new rubric contained the following three domains: Content, Language and Effectiveness, and Layout and Organisation. These three domains were divided into three sub domains, and each was scored on a 4-point scale. The scale ranged from zero to three, therefore, reducing the possibility of centralised rating. The scale ranged from zero, meaning insufficient, to one, meaning just below target level, to two, meaning on target level, and to three, above target level.

A final alteration of the original SLO rubric originated from the participating teachers, who specifically requested the inclusion of three sub domains; the first was scores on spelling and grammar and the second and third were specific genre descriptors for Layout and Organisation elements for introduction and conclusion of the composition. Data was collected on raters' ratings to analyse for rater effects and G-coefficients.

Table 5.3

Designed Rating Rubrics containing descriptors for 3 Domains with 3 sub domains each, each scored on a 4-point scale ranging from 0 to 3.

	0	1	2	3	Score
	The execution of the assignment is not logically structured and/or includes irrelevant choices.	The execution of the assignment is logically structured but lacks relevant steps/choices.	The execution of the assignment is logically structured, but the steps do not fully connect logically to provide a complete train of thought.	The execution of the assignment is logically structured and includes sufficient relevant steps that clarify the message for the reader.	
Content	The chosen text structure does not align with the theme/topic.	The chosen text structure aligns with the main question but is not fully developed: examples and perspectives are not sufficiently elaborated to fit the theme/topic.	The chosen text structure aligns with the main question: examples and perspectives are sufficiently elaborated to fit the theme/topic.	The chosen text structure supports and clarifies the main question: perspectives and examples are elaborated such that they logically follow from the main question and contribute to the final conclusion.	
	No sources are used.	The message of the text is supported by sources, but the sources do not add substantial content to the message.	The message of the text is supported by a limited number of sources.	The message of the text is substantially supported by multiple sources. Together they form a coherent whole.	
	Language use is not appropriate for the purpose and audience.	Language use is somewhat tailored to the purpose and audience.	Language use is sufficiently tailored to the purpose and audience.	Language use is applied in such a way that it does not raise questions regarding the purpose and audience.	
Language & Effectiveness	The text contains so many grammatical, punctuation, and spelling errors that the message is confusing.	The text contains a significant number of grammatical, punctuation, and spelling errors, making the message unconvincing.	The text contains some grammatical, punctuation, and spelling errors that do not hinder the conveyance of the message in the text.	The text is nearly flawless, and the use of spelling, punctuation, and grammar clarifies and strengthens the message of the text.	
	No mechanisms, such as imagery and stylistic devices, are used to engage the reader.	The text contains some standard mechanisms, such as imagery and stylistic devices, to engage the reader.	The text contains the desired mechanisms, such as imagery and stylistic devices, to engage the reader.	The text shows deliberate and excellent use of mechanisms, such as imagery and stylistic devices, to engage the reader.	
	The text is not divided into paragraphs. Paragraphs, sentences, and clauses are not logically connected, making the message unclear.	The text is not divided into paragraphs. Paragraphs, sentences, and clauses are not logically connected, making the message unconvincing.	The text is divided into meaningful paragraphs. An attempt is made to connect paragraphs, clauses, and sentences, making the message convincing.	The text is structured such that meaningful paragraphs, sentences, and words logically follow one another and strengthen the message.	
Layout & Organisation	Elements for clarifying the introduction, such as introduction of the topic, question, and subject, are not present.	A single element for clarifying the introduction, such as the introduction of the topic, question, and subject, is present.	Multiple elements for clarifying the introduction, such as introduction of the topic, subject, and question, are present.	Elements for clarifying the introduction, such as the introduction of the topic, subject, and question, are used excellently to strengthen the structure of the text.	
	Elements for clarifying the conclusion, such as paraphrasing the question, summary of perspectives, and clincher, are not present.	A single element for clarifying the conclusion, such as paraphrasing the question, summary of perspectives, or clincher, is present.	Multiple elements for clarifying the conclusion, such as paraphrasing the question, summary of perspectives, and clincher, are present.	Multiple elements for clarifying the conclusion, such as paraphrasing the question, summary of perspectives, and clincher, are used excellently to strengthen the structure of the text.	
				Final Score	

5.2.2 Teacher Interview

Teachers were interviewed about their experiences of teaching of prerequisite writing skills and rating of the assessment after rating was completed. The collected data was expected to contribute to an overall validity argument, in particular backing for warrants for evaluation inferences for the rating procedure (Knoch & Chapelle, 2018). The data was collected through a semi-structured interview with both teachers, which lasted about 45 minutes. The interview addressed five topics; teaching materials, and teachers' domain and pedagogical knowledge, rating procedures and time needed for rating.

The rater interview was recorded with the *iPhone Dictaphone* App. It was transcribed full verbatim into segments based on speaking turns of the interviewer and the respective interviewees. The interview transcript was taken through three cycles of coding to arrive at a coding scheme based on 2 assumptions for the evaluation inference as proposed by Knoch & Chapelle (2018), in which evaluation is defined as "observations are evaluated using procedures that provide observed scores with intended characteristics" (Knoch & Chapelle, 2018, pp483) These assumptions identify whether a) raters were able to identify differences in performance levels across score levels, and b) raters were comfortable using the scales in the rating rubrics. The transcript was coded by one coder using sentence-based coding, based on separate sentences with a single communicative function.

5.2.3 Rater Questionnaire

Qualitative data about external raters' experiences of the rating process were collected through an open-ended rater questionnaire after the rating procedures. Raters' beliefs about writing may affect writing scores. Therefore, beliefs about writing were analysed. All raters were asked what aspects of writing they believed were most important and which they emphasized in their lessons. Examples of questions in the open-ended questionnaire were "*What do you consider effective in your writing classes? What do you believe is significant in teaching writing? How did you execute the rating procedure?; Did you alter scores already assigned? Why?"* The full set of questions can be found in appendix A.

Thematic analysis on time- intensity of the rating procedure, the rating procedure itself, rater beliefs, and rater perceptions of the rating form was carried out through coding based on the above-mentioned assumptions for the evaluation inference by Knoch & Chapelle (2018).

5.2.4 Student Questionnaires

An online *Qualtrics* questionnaire administered to students investigated student perceptions on the three aspects of assessment accessibility, 1) the opportunity to learn, 2) academic enablers, which means that students have the attitudes and engagement behaviours that enable them to participate in the assessment without barriers and 3) assessment accessibility in that the assessment itself was structured, presented, and administered in such a way that students can manifest their writing skills without experiencing hindrances (Elliott et al., 2018).

Items in the questionnaire inquired after student perceptions of the writing lessons; their experienced support and hindrances during assessment administration; appendix B shows the complete student perception questionnaire.

Opportunity to Learn. Opportunity to learn was measured in five-point Likert items inquiring after how much students perceived to have learned from for example ‘naming text genres and purposes of sources in the documentation file’ or ‘receiving teacher feedback’. Students were enquired after their experience of applying these strategies by the same items in the questionnaire immediately after the assessment, asking them about how stressful they perceived applying these same elements.

Academic Enablers. Academic enablers inquired after attitudes and engagement behaviours that make it easier for students to carry out an assessment task, for example a) engagement and b) motivation and anxiety (Elliott et al., 2018).

Engagement was measured through items in the student questionnaire that inquired after the time spent on familiarisation with the documentation file and the number of additional sources students looked for. Items also inquired after the perceived difficulty level and usefulness of sources in the documentation folder and the ease with which students

could carry out assignments for which they needed the previously instructed writing strategies. Items were answered on a 5-point Likert scale.

Attitude to writing was measured by employing the Writing Apprehension Test (Daly & Miller, 1975). Students were administered an adapted version of the Writing Apprehension Test in an online Microsoft Forms environment at the beginning of the semester. Two statements were removed from the questionnaire and final scores were adapted, and following the rating procedures of the authors, the scale moved accordingly. Statements in the test were, for example, *I avoid writing,* and *I am afraid of writing essays when I know they will be evaluated* (Appendix C). Students were asked to agree or disagree on a five-point Likert scale.

Assessment Accessibility. Items in the questionnaire inquiring into students' perceptions of assessment accessibility asked about the assessment environment, the clarity of the assessment task, their understanding of the task and clarity about standardised factors during assessment administration, such as use of devices, possibility to ask questions.

5.3 Data Collection Procedures

The Writing Apprehension Test was administered to students at the start of the school year in 2023 through an Online Microsoft Forms questionnaire. The data was coded and cleaned in Microsoft Excel. Also, the two participating teachers received the lesson series and the documentation file at the start of the semester.

Two weeks before the assessment was administered, in October 2023, teachers were sent the task and rating rubrics. In preparation, students wrote a skeletal outline of 200 words maximum including references. During assessment administration, in the first week of November 2023, which lasted 120 minutes and in the presence of their own teachers, students wrote an expository essay on their own devices in exam mode, meaning that students did not have access to internet or their own files. Although, they were not allowed to use spelling and grammar checks, they were allowed to use their pre-constructed skeletal outline. Those students who had been selected for the rating sample but were absent during

assessment administration were substituted with students from the same grade based on shared similarity in their relative position towards classmates' performance in Dutch language. Immediately after assessment administration, all students filled out the questionnaire.

Teacher raters were asked to assign scores for each of the nine sub domains on the rating form for each student's writing performance. They were instructed not to confer, and to assign scores individually. They also functioned as second rater for selected students who were not in their classes. This means that all students were rated by their own teacher, and selected students were rated twice by the second teacher from the same school. After initial scoring the researcher had an interview with both teachers, which took place in the first week of December 2023 at the school.

At the end of November, four external raters were sent 26 anonymized student performances, the assessment task and documentation file along with the rating rubric through email. Raters were instructed to assign scores for each of the nine sub domains in the rating form for each of the 26 performances, similar to how they would rate their own students. They scored at home in their own time without supervision and returned the assigned scores through email. Raters were not informed about students' grading history, or that performances originated from students taught by different teachers. Nor did external raters receive training or writing exemplars for grading. Raters also received an open-ended questionnaire. It inquired after their perception of the rating process and their interpretation of criteria. Raters were asked to return the completed rating forms and filled out questionnaires by email within two weeks.

All raters were told that the rating process would take approximately 8 hours. All raters received a compensation per rated performance similar to compensation of second raters during regular Dutch exam ratings.

5.4 Data Analysis

Data analysis was divided into two parts. The first part answered the question to what extent raters rate consistently by obtaining G-coefficients from quantitative rater data and

exploring qualitative rater data. The second part of data analysis investigated students' perceptions of assessment accessibility and how these are related to the obtained writing scores.

5.4.1 Quantitative Rater Data Analysis

Quantitative rater data was collected through rating forms and coded into a dataset which was analysed with *SPSS26*. Missing data were removed from the dataset (n=26). Six raters observed three domains divided over nine sub domains.

Descriptive statistics for all domains per rater were obtained and analysed to find support for the evaluation inference assumption raters can consistently apply the scale to test task (Knoch and Chappelle, 2018).

Also, generalizability theory (Webb & Shavelson, 2005) treats student and rater as facets in a variance analysis. A components analysis estimates variances in student and rater scores and their interaction effects. In this way, the amount of measurement error both student and rater contribute to the observed score can be estimated, as well as the assessment's reliability. In rater studies, the G-coefficient explains consistency among raters by estimating the relative score, whereas the absolute G-coefficient explains consistency of individual raters as well as their individual validity of rating.

In this study, the G-coefficient for absolute agreement refers to consistency of individual raters about the exact level of the scores, whereas relative agreement is about the consistency in the ordering of students' score levels, regardless of the actual precise scores. Multiple univariate analyses with a factor student on the rating sub domains were run to obtain variance components for the absolute scores. A second univariate analysis with factor p x r on all rating sub domains was run to obtain variance components for relative scores.

The calculated variance components were then copied into an Excel formula sheet and a G coefficient was calculated with the following formula

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + (\hat{\sigma}_r^2 + \hat{\sigma}_{res}^2)/K}$$

in which

p = student performance

r = rater

res = residual score

k = total number of observations by raters.

Rater reliability was estimated through the following formula

$$\hat{\rho}^2 = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_r^2 + \hat{\sigma}_{res}^2 / K}$$

in which

p = student performance

r = rater

res = residual score

k = total number of observations by raters.

5.4.2 Qualitative Rater Data Analysis

Qualitative rater data was collected through the rater interview with teacher raters and the rater questionnaire for external raters.

Teacher Interview. The transcript was thematically analysed, aimed at finding support for the assumptions that raters could rate reliably at task level. The transcript was analysed to find support for a) raters being able to distinguish between differences in performances, b) whether raters were comfortable when assigning scores to scales. This was done by coding and focussing on the following themes: rater preparedness for rating, rating method, wording in rating sub domains, interpretation of rating sub domains, raters' beliefs about writing and finally comparability of student scores.

Rater Questionnaire. Rater questionnaires were analysed to obtain data for raters' experiences, such as preferences and beliefs about teaching writing to explain possible biases in rating. Also, qualitative data were collected on raters' perceptions to find support for a) raters being able to distinguish between differences in performances, b) whether raters were comfortable when assigning scores to scales.

5.4.3 Student Quantitative Data Analysis

The student data set consisted of responses to a questionnaire inquiring after assessment accessibility, including the results of the Writing Apprehension Test and students'

writing assessment scores. Following the test manual instructions (Daly & Miller, 1975) the raw scores from the Writing Apprehension Test were recoded into three relevant domains; those with *writing anxiety (*scores 11-44), *neutral* attitude to writing (scores 45-86) and students with *a lack of motivation for writing* (scores 87-120). Students' writing anxiety final scores were added to the student questionnaire data.

Descriptive statistics and correlation analyses were executed with *SPSS26* to find out to what extent students' perceived assessment accessibility was correlated to their obtained writing scores.

6 Results

6.1 Rater Agreement and Evaluation of Rating Processes

Table 6.1 shows the mean ratings and standard deviations per rater for each of the three writing domains and the final score for students' writing performances. The maximum score of the assessment was 27 points, with a maximum score of nine for each domain.

The domain Content shows a mean of 5.91 with a minimum average of 5.11 for rater

Table 6.1

Assigned Mean Scores and Standard Deviations for Domain Totals and Final Scores per Rater for Students' Writing Performances

		Content		Language & Effectiveness		Layout & Organisation		Final Score	
		M	SD	M	SD	M	SD	M	SD
Teacher raters	Rater 1	5.81	1.33	4.04	1.37	5.89	1.58	15.74	3.22
	Rater 2	5.11	1.61	3.57	0.95	5.46	1.36	14.15	2.62
Overall Teacher Raters		5.50	1.50	3.85	1.18	5.69	1.49	15.03	3.00
External raters	Rater 3	6.62	1.90	3.77	1.82	6.04	1.54	16.42	4.72
	Rater 4	5.61	0.94	4.35*	1.32	5.23	0.65	15.19	1.98
	Rater 5	6.27	1.54	3.65	1.13	5.88	1.11	15.81	2.93
	Rater 6	6.08	1.29	2.69	1.29	5.38	0.90	14.15	2.66
Overall External Raters		6.14	1.48	3.62	1.52	5.63	1.13	15.39	3.30
Overall		5.91	1.51	3.68	1.41	5.65	1.26	15.25	3.20

Note. Based on ratings of 26 performances per rater.

* rater 4 ratings of sub domain Language and Effectiveness: Use of literary devices and imagery to engage reader were removed from the analysis due to 0 scores for all students.

2 and a maximum average of 6.62 for rater 3. The domain Layout and Organisation has a mean score of 5.65 with a minimum of 5.38 for rater 6 and a maximum of 6.04 for rater 3. Whereas the domains Content and Layout and Organisation just above the middle of the 9-point scale, the domain Language & Effectiveness shows a lower mean of 3.68, with rater 6 mean average ratings of 2.69 and rater 1 with a mean score 4.04.

6.1.1 Rater Agreement

The G-coefficient for the total score, using teacher raters amounts to .82 for absolute agreement. This means that when six random raters score the assessment task 82% of the variance in the rating is due to student differences in performance, while 18% is due to measurement error. The G-coefficient of .81 for relative agreement indicates that the variance in scores is consistent across the six raters and acceptable for high-stakes writing assessment.

Additionally, G-coefficients were calculated for teacher raters and for external raters separately. The results can be seen in table 6.2. Although the G-coefficients drop to .68 and .70 respectively for teacher and external raters.

For teacher raters the relative G-coefficient is only slightly lower than the absolute G-coefficient.

Table 6.2

G-coefficients for Absolute and Relative Agreement for all Six Raters, Two Teacher Raters, and Four External Raters regarding Final Score of Writing Assessment.*

	G-coefficient absolute agreement	G-coefficient relative agreement (Reliability)
All raters (r=6)	.82	.81
Teacher raters (r=2)	.70	.68
External raters (r=4)	.71	.70

*Note** = based on 26 scored writing performances.

6.1.2 Rater Experiences

Qualitative data collection from raters focused on the experiences of all six raters were meant to investigate whether three scoring inference assumptions were fulfilled; a)

raters were able to identify differences in performance levels across score levels, and b) raters were comfortable using the scales in the rating rubrics.

Rater Preparedness for Rating. Unlike practice in other countries, raters in The Netherlands are not specifically trained. It is, therefore, essential to find out how Dutch raters, teachers, or external raters, prepare themselves for their rating task.

Both teacher raters and external raters were given a time span of two weeks to complete rating. They were not monitored during rating. Two of the external raters returned the scored performances within a week, the other raters, including the teacher raters, took the full two weeks to complete the rating task. All raters were qualified teachers of Dutch with at least 2 years of experience teaching exam years as can be read in table 5.1.

Teacher raters indicated to have read each performance twice before assigning ratings, which, according to the teacher raters, is comparable to rating a regular school-based writing assessment. For the rating of their colleagues' students, however, they sufficed with only one reading of the performance. The difference in rating for a colleague's students lies in the fact that they, as teachers, do not have to provide students feedback on the rated performance. Timewise, teacher raters reported to have spent on average 20 minutes per performance and checked all performances two more times, resulting in a total of nine hours. External raters, on the other hand, indicated to have spent an average of eight hours. The two teacher raters reported to have needed preparation time to be able to familiarise themselves with the writing task, the rating form and subdomain score assignment. Training, exemplars, and benchmarks were not provided for any of the six raters, however, raters did not report to have perceived this as a hindrance, except one. This rater would have liked to have received exemplars to compare her ratings with an exemplar.

Rating Method. Even though the rating rubrics were standardised, raters' various preferred methods of rating could affect scores. Lumley (2002) sets out three stages of rating performance tasks: 1) reading the performed task, 2) rating the sub domains, 3) consideration of scores given in stage 2. Although raters reported having used these three stages; three external raters indicated to have thoroughly studied the assessment task and

afterwards scored each performance for all sub domains. External rater 6 reported a similar procedure but reported to have reconsidered performances with similar scores in a third round. Both rater 3 and 5 reconsidered the first five performances they scored after they had finished rating the complete set of performances. Rater 5 reported to have evaluated uncertainties about score assignment in a second stage.

Wording in Rating Domains. Wording of the sub domain Grammar, Interpunction, and Spelling Mistakes and the sub domain References were perceived as ambiguous by rater 2, 3 and 6, for quantifiers such as ‘some’ and ‘several’ should have been made concrete. Additionally, raters reported to believe that the Grammar, Interpunction, and Spelling Mistakes sub domain was too lenient in wording, because a text “rife with spelling errors communicates the message, but still does not meet requirements.”

Interpretation of Rating Domains. Raters 5 and 6 reported to have perceived an overlap between the sub domains in Content and Layout and Organisation. For if a student does not do well in the domain content, they might also not receive scores for another sub domain in that domain, such as paragraphing. Another issue for the sub domains Elements that clarify introduction or conclusion are used was mentioned; students may have used more than two elements that were not executed well, but still earn points.

Also, raters pointed out that there may have been overlap in sub domains Execution of Task and Effective Use of Language for Audience and Purpose. Raters reported to believe that Execution of Task already required students to focus on audience and purpose.

Interestingly, rater 2 employed half scale points on her rating forms, thereby increasing the scale range and differentiating between students’ abilities on a similar scale level. This indicates that, that at least for this rater, a longer scale could serve as even more discriminating for students’ abilities.

Rater’s Beliefs about Writing. Raters’ beliefs about writing may affect writing scores. All raters were asked what aspects of writing they believed were most important and which they emphasized in their lessons. Raters share the belief that spelling, use of language, and cohesion are key domains in writing. Two of the six raters also consider being able to make

references a principal domain. However, teaching writing conventions for specific genres is not mentioned as an important strategy for writing by any of the raters. Moreover, teacher raters indicated that certain writing domains were not explicitly taught, such as Referencing, Rhetorical Devices, and Structure Words. Therefore, teacher raters felt they were lenient in these domains.

Table 6.3

Summary of Supportive and Hindering Elements Raters Experienced before and during the Rating Process

	Hindrance	*Support*
Rater Preparedness		
Rater Preparedness	Lack of calibration set mentioned by one teacher rater.	Experience teaching exam years.
Raters are comfortable when applying descriptors and confident in their decisions		
Wording in Rubric	Abstract wording in certain sub domains may cause inconsistencies in rating process.	
Interpretation of Rubric	Overlap between certain rating sub domains.	
Raters are able to identify differences in performances across score levels		
Interpretation of Rubric	Scale length may need extension.	
Raters can consistently apply the scale on test task		
Rating Method		Routine method of rating (stages of rating) can be employed. Time required for rating is like current practices.
Rater Beliefs about Writing	*Spelling*, *use of language*, and *cohesion* are key domains in writing. Rating sub domains leniently when not taught.	
Comparability of Scores		Standardising the assessment improves comparability of students' writing scores.

Comparability of Scores. Teacher raters indicated to have found comparing students' writing abilities on the same writing topic easier than comparing students' scores on different writing topics. In previous school-based writing exams, teachers would design

similar writing assessment tasks for four to five different content topics and have students build their own documentation files. Teacher raters also indicated to have gained better insights into student achievement in certain domains. The rubrics supported teachers in identifying on what domains of writing their students performed well and which domains students needed improve in.

The results of the findings of the interview and survey are summarised in table 6.3.

6.2 Student Perceptions: Opportunity to Learn, Academic Enablers, and Assessment Accessibility

Students were questioned about their perceptions of assessment accessibility. Three elements that might affect assessment accessibility are reported below: 1) students' opportunities to learn, 2) their access skills and motivation, and 3) their perceptions of the assessment task and administration setting.

6.2.1 Students' Perceptions: Opportunity to Learn

Students were questioned about their perceived effectiveness of classroom instructed writing strategies. These strategies were uniformalised for all students and were instructed by teachers. Students tended to answer in the centre of the scale. Students perceived receiving

Table 6.4

Descriptive Statistics for Students' Perceived Effectiveness of Instructed Writing Strategies

Perceived Effectiveness of Taught Writing Strategies	N	M	SD	Range
Identifying text purpose and genres of sources	68	2.43	0.95	0-5
Ranking Chat GPT essays	68	2.56	1.19	0-5
Creating a skeletal outline	68	3.12	1.18	0-5
Giving peerfeedback	68	2.32	1.32	0-5
Receiving teacher feedback	68	2.19	1.67	0-5
Rating written compositions with rating rubrics	68	2.49	1.47	0-5
Rewriting composition after receiving feedback	68	2.40	1.47	0-5
Explicit instruction on expository writing	68	3.09	1.28	0-5
Total score for preparedness (k=8, α =.72)	68	2.57	0.76	0-5

Note. Scale in which 0 means 'learnt nothing' and 5 means 'learnt a lot'.

teacher feedback as somewhat effective (M=2.19), on the other hand, explicit instruction by teachers (M= 3.09) and creating a skeletal outline (M=3.12) were perceived as instructed writing strategies that students moderately learned from. Reliability analysis shows that the

scale variable Perceived Effectiveness of Classroom Instructed Writing Strategies has an internal consistency of .72. Hence, an aggregate scale, which can be seen in table 5.4, for Perceived Effectiveness of Instructed Writing Strategies was used for subsequent correlation analysis.

Students were also asked to rate their stress experience when applying the various instructed writing strategies, which can be seen in Table 6.5. On average, students rated their perceived stress for applying all instructed writing strategies between M= 1.13 and M=1.74, apart from the strategy creating a skeletal outline (M=2.59), which on average, students rated as more stressful than applying the other instructed writing strategies.

Reliability analysis results show that the scale variable perceived stress experience of instructed writing strategies has an internal consistency of .89. Thus, an aggregate scale for perceived stress experience applying instructed writing strategies (M= 1.55) was used for analysis to check for a correlation with writing achievement.

Table 6.5

Descriptive Statistics for Student Experiences of Applying the Instructed Writing Strategies

Stress Experience Preparation	N	M	SD	Range
Identifying text purpose and genres	68	1.74	1.19	0-5
Ranking Chat GPT essays	68	1.25	1.18	0-5
Creating skeletal outline	68	2.59	1.42	0-5
Giving peer feedback	68	1.13	1.09	0-5
Receiving teacher feedback	68	1.43	1.44	0-5
Rating with rating rubrics	68	1.34	1.27	0-5
Rewriting after receiving feedback	68	1.62	1.39	0-5
Explicit instruction on expository writing	68	1.29	1.40	0-5
Experience Writing Preparations (k=8, α= .89)	68	1.55	0.98	0-5

Note. On a scale in which 0 means ‘not stressful at all’ and 5 is ‘very stressful’.

6.2.2 Students’ Perceptions: Academic Enablers

Academic enablers, such as engagement behaviours and attitude were measured through preparation before the assessment administration and attitude towards writing. Table 6.6 shows that students spent between one and nine hours with a mean of 3.13 hours familiarising themselves with the sources in the documentation file. Students looked for

between one and seven additional sources with an average of almost two additional sources (M=1.92) in preparation for assessment administration.

Table 6.6

Descriptive Statistics for Individual Preparation Activities by Students before the Assessment Administration

	N	M	SD	Range
Familiarisation with documentation file in hours	68	3.13	1.84	1-9
Number of additional sources looked for per source	60	1.92	1.77	0-7

Note. On a continuous scale from 0.

Other items for measuring engagement included items inquiring after the perceived ease with which students could create a skeletal outline for the assessment task (M= 2.99), and write a mock expository essay (M=3.35). The results are shown in table 6.7. It shows that students tended to respond in the centre of the scale. Reliability analysis show that the scale variable Perceived Ease of Preparatory Activities for Writing Assessment has an internal consistency of .68. Hence, an aggregated scale for perceived ease of preparatory activities for writing assessment can be used for correlation analysis.

Table 6.7

Descriptive Statistics for Students' Perceived Ease of Preparatory Activities for Writing Assessment

	N	M	SD	Range
Creating skeletal outline individually	68	2.99	0.87	1-5
Writing a mock expository essay	68	3.35	0.81	1-5
Ease of Preparatory Activities for Assessment (k=2, α =.68)	68	3.17	0.73	1-5

Note. Scale in which 1 is 'very easy' and 5 is 'very difficult'.

A final aspect inquiring after students' engagement behaviours as academic enabler were items inquiring after students' perceived difficulty level and usability of sources in the documentation file. The results can be seen in table 6.8. Students perceived the difficulty level of the sources in the documentation file as neither difficult nor easy (M=2.84). Usability of the sources in the documentation file is perceived similarly, as neither useful nor useless (M=2.54).

Table 6.8

Descriptive Statistics and Reliability Analysis for Student Perceptions of Difficulty Level and Usability of Provided Support Materials

	N	M	SD	Range
Perceived difficulty level sources in documentation file	68	2.84	0.54	1 - 5
Usability of sources in documentation file	68	2.54	1.03	1 - 5
Difficulty level and Usability of Provided Support Materials (k=2, α= .52)	68	2.69	0.66	1 - 5

Note. In which the Likert scale runs from 1 perceived as easy to 5 being very difficult.

Reliability analysis results show that the scale variable *perceived difficulty level and usability of provided support materials* has an internal consistency of .52, which is not acceptable for constructing an aggregate scale. Therefore, a correlation analysis per individual item and obtained scores was calculated.

Motivation was measured through the Writing Apprehension Test (Daly & Miller, 1975). Students were categorized in three groups, the ones with writing anxiety (n=2), the ones that lack motivation to improve writing (n=12) and those that have neither writing anxiety nor lack motivation for writing (n=41) (Daly & Miller, 1975).

6.2.3 Students' Perceptions: Assessment Accessibility

Table 6.9 shows the factors of Students' Experienced Feelings during Assessment Administration. Over the whole, students indicated to have felt only slightly hindered; experience blackout M=1.37, experiencing lack of time to express ideas M=1.28, test anxiety

Table 6.9

Descriptive Statistics and Reliability Analysis for Students' Experienced Feelings during Assessment Administration

Experience	N	M	SD	Range
blackout	68	1.37	0.62	1-4
lack of time to express ideas	67	1.28	0.74	1-4
test anxiety	64	2.08	0.86	1-4
issues due to insufficient preparation	65	1.58	0.71	1-4
issues due to ambiguity in assessment task description	67	1.31	0.66	1-4
a lack of ideas for performing the task	68	1.51	0.74	1-4
Experienced Feelings Assessment Administration (k=6, α= .59)	59	1.52	0.42	1-4

Note. On scale in which 1 means 'not hindered at all' and 4 means 'extremely hindered'.

M=2.08, issues due to insufficient preparation M=1.58, experiencing issues due to ambiguity in assessment task description M=1.31, experiencing a lack of ideas for performing the task M=1.51. Experiencing test anxiety stands out in *experienced issues during assessment administration*.

Reliability analysis results show that the scale variable Students' Experienced Feelings during Assessment Administration has an internal consistency of .59.

Administration Assessment Setting was measured through four items as shown in table 6.10. It enquired after students' experiences of the tranquillity of the environment (M=3.75), the opportunity to ask questions (M=3.88), the digital environment (M=4.01) and the clarity on permitted aids (M=4.00). Reliability analysis of these four items presented an internal consistency of .81. Hence, an aggregate scale for Students' Perceptions of the Administration Setting can be used for subsequent correlation analysis with obtained writing scores.

Table 6.10

Descriptive Statistics and Reliability Analysis for Students' Perceptions of Administration Setting

Administration setting	N	M	SD	Range
was calm and quiet	68	3.75	0.95	1-5
offered opportunity to ask questions	68	3.88	0.91	1-5
provided optimal working devices	68	4.01	0.72	1-5
provided clarity about permitted aids	68	4.00	0.90	1-5
Total Score Perception Administration Setting (k=4, α= .81)	68	3.91	0.69	1 -5

Note. On a scale in which 1 means 'totally disagree' and 5 means 'totally agree'.

Other items that measured assessment accessibility are in table **6.11**. Thus, factors such as support from a pre-drafted skeletal outline (M=1.01) and allotted time (M=2.19) were evaluated positively by students.

Table 6.11

Descriptive Statistics for Usability of Aids during Assessment Administration

Aids during Assessment Administration	N	M	SD	Range
Support from skeletal outline	68	2.19	1.18	1-5
Time slot administration	68	1.01	0.12	1-2

Note. On a scale in which 1 means a lot and 5 means not at all.

6.2.4 Correlations between Perceived Assessment Accessibility and Obtained Writing Scores

Table 6.12 shows the correlations between students' perceptions of the three elements of assessment accessibility and obtained writing scores. It shows that perceived effectiveness of instructed writing strategies (.258), perceived difficulty of sources in documentation file (-.289) and administration setting (.278) significantly correlated with the obtained writing scores. A weak, significant, negative relationship is found between perceived difficulty of sources in documentation file and obtained writing scores (-.289) and a weak, significant, positive relationship between effectiveness of instructed writing strategies and obtained writing scores (.258) and a small, significant, positive correlation between administration setting and obtained writing scores (.278).

In contrast to the Writing Anxiety Test, which identified merely 2 students with anxiety, 18 students responded to have been 'a little' to 'extremely hindered' by test anxiety during the assessment administration in the survey. Thus, experiencing anxiety as a hindrance

Table 6.12

Descriptive Statistics and Correlation Table for Elements of Assessment Accessibility and Obtained Writing Scores

	N	M	SD	Obtained Writing Score
Opportunity to Learn				
Perceived effectiveness of writing strategies	64	2.57	0.76	.258*
Perceived experience applying instructed writing strategies	64	1.55	0.98	.029
Academic Enablers				
Familiarisation with documentation file in hours	64	3.13	1.84	-.023
Number of additional sources looked for per source	57	1.92	1.77	-.108
Perceived ease of preparatory activities for writing assessment	64	2.69	0.66	-.132
Perceived difficulty level of sources in documentation file	64	2.99	0.87	-.289*
Usability of sources in documentation file	64	3.35	0.81	-.210
Assessment Accessibility				
Administration assessment setting	64	3.91	0.69	.278*
Support skeletal outline during administration	64	2.19	1.18	.017
Timeslot administration	64	1.01	0.12	.025
Students' experienced feelings during assessment administration	64	1.52	0.42	-.116

*. Correlation is significant at the .05 level (2-tailed).

stood out in Experienced Issues during Assessment Administration and a variance analysis was conducted to examine differences in Obtained Writing Scores between groups of

students with anxiety and students without anxiety. No significant differences were found between the groups, F(2,58) =1.049, *p* = .357 at 0.05 level.

Figures 6.1, 6.2, and 6.3 show the scatterplots for the aggregate scales that significantly correlate on obtained writing scores.

The scatterplot in figure 6.1 shows a weak, positive relationship between Perceived Effectiveness of Classroom Instructed Writing Strategies and Obtained Writing Scores. The scatterplot in figure 6.2 shows the significant relation between the perceived difficulty and obtained scores for the writing assessment and figure 6.3 shows the scatterplot for the

Figure 6.1

Scatterplot for Students' Perceived Effectiveness of Instructed Writing Strategies on Total Writing Score

Figure 6.2

Scatterplot of Students' Perceived Difficulty of the Documentation File on Obtained Writing Grade

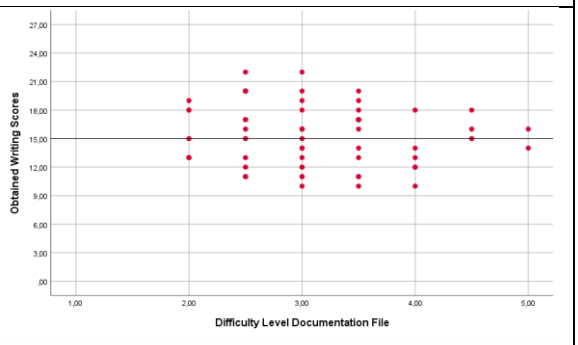
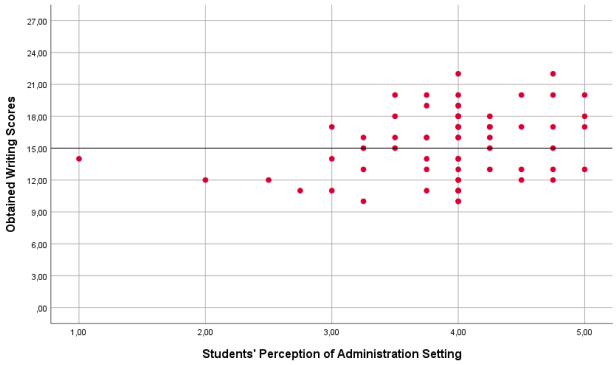


Figure 6.3

Scatterplot for Student Perceived Assessment Administration Setting Total Score on Total Writing Score



relation between Students' Experienced Feelings during Assessment Administration and Total Writing Score.

7 Discussion

This study investigated if and how it is possible to re-introduce a centralised writing exam for Dutch and was carried out for the Dutch Institute for Educational Measurement. It showed that a newly designed, high-stakes, documented writing task was rated consistently by two teacher raters without training or instruction. In addition, it showed that teacher raters are able to discriminate between students' abilities and share a mutual understanding of the construct. Moreover, the study found that students perceived high assessment accessibility in terms of opportunities to learn, preparatory assignments, and supportive administration setting. Practical implications for the re-introduction of a centralised writing exam are discussed later on.

7.1 Conclusions

Rater Agreement

- *To what extent do raters consistently assign scores to performers on a documented writing task administered to students in their final year of pre-university education in a high-stakes assessment setting, using a rating procedure designed for this purpose?*

Rater agreement estimates were acceptable to good with G-coefficients of .81 and .84 for absolute and relative agreement using six untrained raters. This is not in line with previous research which suggests that G coefficients of .80 are exceptions (Hamp-Lyons, 2012). However, it is in line with Skar & Jølle (2017), who find a .93 ICC based on ratings of narrative and expository texts by 8 trained teachers.

In terms of validation of the writing assessment, the warrants of different raters assigning the same ratings to responses are backed. Although raters seemed to rate more severely in the domain language and effectiveness than the domains content, and layout and organisation, and a one way Anova shows that there are significant differences in strictness between the raters at 0.05 level, $F(5,156) = 2.295$, $p = .048$, a post-hoc Scheffé test does not show a significant deviation in scores for one single rater compared to the other raters at

p=.05. Therefore, the evaluation inference assumption: that raters are able to consistently apply the scale to the assessment task is fulfilled. Hence, raters rate consistently at test level and the number of raters is sufficient to arrive at a reliable score (Knoch & Chapelle, 2018).

- *To what extent do teacher raters' scores of a high-stakes writing assessment show rater agreement?*

Although rater agreement estimates for six raters are good, in Dutch central examination setting only two raters are used. G- coefficients for two raters were lower, but still promising. Two teacher raters obtained G-coefficients of .70 and .68 for absolute and relative agreement rating a single task for a sample of 26 of their complete population of students. This means that two teacher raters rating independently and without training obtain acceptable rater reliability estimates for a high-stakes writing assessment. The estimates are similar to what was found in previous research. Brown et al. (2004) found adjacent agreement percentages of 70-80% and reliability rates between .70 and .80, after teachers of lower years in secondary education were trained.

- *To what extent do external raters' scores of a high-stakes writing assessment show rater agreement?*

Four raters achieved moderate to good G-coefficients of .71 and .70 for absolute and relative agreement for the sample of 26 performances of the single writing task, which is comparable to teacher raters.

Perceived Supports and Hindrances in Rating Procedures

- *How do raters prepare themselves to rate documented writing assessments?*

Both teacher and external raters familiarise themselves with the task and the rating rubrics before applying stages of rating. Raters use different methods in applying Lumley (2002) three stages of rating, 1) reading the performance, 2) rating the performance and 3) consideration of scores. Some raters entered stage 3, consideration, only for those performances that were similar in scores, after completing a combined reading and rating stage for all performances. Some raters combined reading and rating performances and used reconsideration only for the first five performance they had scored. Yet another rater

indicated to have combined reading and rating and noting down uncertainties, which were revisited in the reconsideration stage.

In preparation to the rating procedures, raters also relied on their experiences as teachers teaching writing and on previous rating of writing assessments. This could be concluded from raters' references to the central examinations' guidelines regarding use of spelling and grammar and to holistic and analytic rubrics that had been used previously.

- *Which aspects of the assessment procedure for the documented writing assessment support or hinder raters in arriving at valid interpretations of writing products and deriving a score based on these interpretations?*

Both teacher and external raters indicated to have felt hindered by wording in rubrics as well as their individual interpretations of rubrics, it caused them to feel unsure about assigning scores to the domains. Teachers indicated that assigning discriminating scores was based on their experiences with rating and familiarity with students. It seems that rater interpretation of rubric and wording can be considered a hindrance, however, found rater agreement levels show that this group of six raters, without training, were consistent in their beliefs about the concept documented writing and rating behaviours. It appears that the raters, all experienced teachers, involved in this study have a shared understanding of the concept documented writing and its domains and the hindrances they perceived may have been due to working with an unfamiliar rating rubric.

Student Perceptions of Writing Assessment Accessibility

- *To what extent do students perceive the documented writing assessment procedure accessible?*

This part of the study was explored by looking at students' perceived opportunities to learn, academic enablers in terms of engagement behaviours and attitude towards writing, and perceived assessment accessibility during assessment administration.

Opportunity to Learn

Students' perceptions of opportunities to learn were measured through their evaluation of the effectiveness of the instructed writing strategies and the ease with which

students felt they could apply the instructed writing strategies. Overall, students indicated that they had learned moderately. They perceived direct teacher instruction and writing a skeletal outline for the assessment task as most valuable. The students' perceived effectiveness of instructed writing strategies showed a positive correlation with the obtained writing scores. There was a non-significant, weak, positive association between experience of applying effective writing strategies and obtained writing scores.

Academic Enablers: Engagement Behaviours and Attitude towards Writing

Students perceived difficulty level of sources in documentation file was significantly, but weakly and negatively associated with obtained writing scores, indicating that the harder students perceived these sources, the lower the obtained writing score.

The Writing Anxiety Test, a validated, self-perception instrument, found 2 students with writing anxiety. In the student questionnaire, 18 students indicated to have felt a little to extremely hindered by test anxiety during assessment administration. A variance analysis showed no significant differences in assessment grades. Therefore, due to a lack of variance no association could be demonstrated between writing anxiety and performance.

Assessment Accessibility during Administration

Students did not perceive obstructions or barriers during assessment administration. Students' experiences during administration, such as not having ideas or experiencing a blackout, did not significantly relate to obtained writing scores. Also, students did not feel hindered by lack of time for completion of the task, and they felt supported by bringing a skeletal outline to the administration. No significant associations were found for any of these with obtained writing scores. Students' perceptions of administration setting, however, did significantly relate to obtained writing scores; there was a small, positive relationship between administration setting and obtained writing scores. However, students indicated not to have felt hindered by administration setting. Over the whole, assessment accessibility during administration was rather high.

To conclude, it appears that students considered themselves sufficiently prepared, they indicated not to have been hindered in their opportunities to learn. Also, students'

attitudes, in terms of anxiety or lack of motivation, did not present obstructions to students manifesting their writing abilities during assessment administration. Also, engagement for writing was moderate, students did not perceive being hindered by their own engagement behaviours. Finally, assessment administration did not present obstacles in students manifesting their true writing abilities.

Thus, to the extent that assessment accessibility affected variability in students' obtained scores, it merely concerned the following factors, effectiveness of the instructed writing strategies (positively), the difficulty level of the sources in the documentation file (negatively) and assessment administration setting (positively).

7.2 Discussion

This study set out to find out how and if writing assessment could be re-introduced in a high-stakes, centralised setting after it was removed from the centralised exams in 1998. It explored to what extent the newly-designed writing task fulfils the conditions of a quality and test that is effective (American Educational Research Association. et al., 2014). Therefore, the validity, reliability, and assessment accessibility of the writing task in this assessment administration were assessed. Table 7.1 presents an overview to what extent these conditions a high-stakes centralised writing assessment task.

For validation of the evaluation inference for rating, this study applied elements of Knoch & Chapelle's (2018) validation framework for rating processes as is described in the conclusion section above.

Regarding reliability, this study found acceptable to high rater agreement for two untrained and uninstructed raters rating independently. This is promising; it means that rater agreement levels are likely to increase to highly acceptable levels for high-stakes assessment, for simple tools, such as detailed rating instructions and anchor texts increase reliability (Humphry & Heldsinger, 2014). However, the high agreement levels found do not fullfill expections drawn up in previous research. Bouwer et al. (2015) stipulated agreement indices of .66 for persuasive and argumentative essays and .80 for personal stories when 5 tasks and 5 raters would be used. A possible cause for the differences in agreement levels

between this study and Bouwer et al.'s study (2015) may be due to the heterogeneity of a primary-school sample compared to the homogeneity of a final-year secondary-school sample. Sampling a larger group of final-year, secondary school students from different schools may show different results in writing ability.

The high levels of agreement in this study also show that teacher knowledge and beliefs about integrated writing are aligned with the construct as it was measured (Jia & Zhang, 2023). However, teachers' beliefs about the domain use of language, such as grammar and spelling show a bias in rating compared to the other two domains. These results are in line with Eckes' (2012) study in which the more important raters believe a domain of the concept they are rating to be, the more severely the domain is rated. By combining qualitative and quantitative rater data on rater beliefs and their ratings, all six raters in this study tend to rate the language and effectiveness domain more strictly. This might be the result of the abstract wording in the sub domains for language and effectiveness. Although, the task and rating rubrics were designed with the utmost care and evidence from theoretical research, misalignment of domain with target group for writing; or even misalignment of rubric and task description could also have been a potential cause for remarkable score for the language & effectiveness domain. A practical implication for the assessment procedure is that the domain use of language needs careful consideration in relation to the complete rubric, for this severe rating may distort writing scores. Practical implications are weighting this domain relative to the other two domains in the rubric. Further research could provide insights on proportional weighting of the domains. Another practical, less costly, solution could be explicit instruction on and / or examples of ratings of the language domain in this task.

The third condition that was explored for qualitative and effective testing, was assessment accessibility. In general, students perceived assessment accessibility as high. Four elements in assessment accessibility are worth noting.

For one, students perceived the instructed writing strategies as positively affecting their writing scores, meaning teachers should consider applying writing instruction centred around these strategies.

Second, unlike other studies (Graham et al., 2007; Sabti et al., 2019), student attitudes and engagement behaviours were not found to be related to the obtained writing scores in this study. This does not mean, however, that there would be no effect from attitudes and engagement behaviours on writing scores. Further research into attitude and engagement behaviours towards writing, measured through validated instruments, should be carried out to determine if and how it affects writing scores in a high-stakes setting.

Third, sources in the documentation file should be assessed on their abstraction level, use of language, and genre by both teachers and assessment designers familiar with the type of student, which is in line with Schoonen's suggestion (2005) as to prevent unwanted variability due to accessibility issues.

A final issue that could possibly affect assessment accessibility is administration setting. In practice, this means considering the circumstances in which students perform the assessment task. Assessment procedures and guidelines for digital environments, and aids and supports during administration should be provided. Also, evaluation after assessment administration by questioning student and/ or teacher experiences could help determine considerations in cut-off scores. Theoretical implications should be explored through the question whether administration setting should be considered a random or systematic factor in determining cut off pass / fail scores.

Table 7.1

Evaluation of Exploration of Re-introduction of Writing as Part of the Centralised Exams for Dutch based on the Conditions Needed for Qualitative and Fair Assessment (American Educational Research Association. et al., 2014)

Validity		
Construct Writing	Processes of writing (planning, translation, and revision) are assessed by inclusion of supportive skeletal outline and documentation file.	✓
	Abstract wording in sub domains for documented writing may need finetuning to prevent misinterpretation by raters.	?
	Weighting of sub domains within the rubric could solve severe rating of certain domains.	?
Rater Beliefs	Raters in this study share mutual understanding of the construct Documented Writing.	✓
	Raters were able to identify differences in student abilities using the provided scoring rubric.	✓
Reliability		
Rater Agreement	Rater agreement is acceptable for this assessment task with the designed scoring rubric and likely to increase with instructions and/ or exemplars.	✓
Rater Beliefs	Raters used the sub domains and the scales in them to rate as intended.	✓
Assessment Accessibility		
Opportunity to Learn	Evidence-based practices for teaching writing, such as the instructed writing strategies, are perceived by students to affect writing scores.	✓
Academic Enablers	The effect of attitude on writing scores, through motivation/ anxiety, should be studied through validated instruments in a larger sample and set off against personal details of students (ie. gender, language disorders).	?
	Engagement behaviours should be triangulated by measuring concrete student action, such as presence during lessons, quality of performed assignments, and participation in lessons.	?
	Difficulty level documentation file as input content knowledge for task needs standardization through protocol checking abstraction and language levels as well as genre by teacher and assessment expert reviewers.	✓
Administration Accessibility	Students believe that assessment administration setting impacts writing scores.	✓
	Further research into assessment administration setting should provide guidelines as to what conditions should be in place for optimal conditions during practical assessment administration.	?

8 Conclusion

This study intended to explore whether a newly-designed, high-stakes, documented writing task could meet the conditions of qualitative and effective assessment. It tried to do so by estimating rater agreement, rater perceptions and assessment accessibility. The relative G-coefficient for six raters based on 26 performances showed the rater agreement levels was high with .82; for two raters relative G-coefficients were acceptable with .68. Further improvement may be attained by offering raters training or exemplars and calibration sets. Qualitative results backed the assumptions for validity warrants for rating and showed that teacher raters share a mutual understanding of the construct documented writing.

Moreover, students were asked for their perceived high assessment accessibility during preparation and assessment administration, if there were any obstacles that could potentially affect test scores, perceived effectiveness of instructed writing strategies and assessment administration setting positively correlated on students' obtained writing scores, and the difficulty level of sources in the documentation file negatively correlated to obtained writing scores. Further research into how these factors could potentially affect specific groups of the student population could affect writing scores.

9 References

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). Standards for Educational and Psychological Testing. (2nd ed.). https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition

Bazerman, C. (2015). What Do Socialcultural Studies of Writing Tell Us about Learning to Write? In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 11–23). The Guildford Press.

Beck, S. W., & Jeffery, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing*, *12*(1), 60–79. https://doi.org/10.1016/j.asw.2007.05.001

Becker, A. (2006). A Review of Writing Model Research Based on Cognitive Processes. In A. Horning & A. Becker (Eds.), *Revision: History, theory, and practice* (pp. 25–49). Parlor Press.

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. (2016). *Bringing Writing research into the classroom the effectiveness of tekster, a newly developed writing program for elementary students*. Interuniversity Centre for Educational Research. https://issuu.com/tekster6/docs/tekster_proefschrift-digi2

Bouwer, R., Koster, M., & van den Bergh, H. (2023). Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. *Assessment in Education: Principles, Policy and Practice*, *30*(3–4), 302–319. https://doi.org/10.1080/0969594X.2023.2241656

Bouwer, R., Van Ockenburg, L., Van Der Loo, J., & Van Weijen, D. (2022). *Literatuurstudie naar de kenmerken van effectief schrijfonderwijs in het voortgezet onderwijs [Literature study into the features of effective writing education in secondary education]*. Utrecht: Universiteit Utrecht. https://www.nro.nl/onderzoeksprojecten/literatuurstudie-naar-kenmerken-van-effectief-schrijfonderwijs-in-het-voortgezet-onderwijs

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Brookhart, S. M. (2013). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy and Practice*, *20*(1), 69–90. https://doi.org/10.1080/0969594X.2012.703170

Brown, G. T. L., Glasswell, K., & Harland, D. (2004a). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*(2), 105–121. https://doi.org/10.1016/J.ASW.2004.07.001

Cahill, R., Kellogg, T., Mertens, A., Turner, C. E., & Whiteford, A. P. (2013). Working Memory in Written Composition: An Evaluation of the 1996 Model. *Journal of Writing Research*, *5*(2), 159–190. https://doi.org/https://doi.org/10.17239/jowr-2013.05.02.1

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–37.

Chan, S., & Yamashita, J. (2022). Integrated writing and its correlates: A meta-analysis. *Assessing Writing*, *54*, 100662. https://doi.org/10.1016/J.ASW.2022.100662

College voor Toetsen en Examens. (2021a). *Nederlands (3F) havo: syllabus centraal examen 2023 [Dutch (3F) havo: syllabus central examination 2023]*. https://www.examenblad.nl/system/files/2021/syllabi/nederlands_3f_havo_versie_2_2023.pdf

College voor Toetsen en Examens. (2021b). *Nederlands (4F) vwo: syllabus centraal examen 2023 [Dutch (4F) vwo: syllabus central examination 2023]*. https://www.examenblad.nl/system/files/2021/syllabi/nederlands_4f_vwo_versie_2_2023.pdf

Curriculum for Norwegian, Pub. L. No. NOR01-06, 1 (2020). https://www.udir.no/lk20/NOR01-06

Daly, J. A., & Miller, M. D. (1975). Writing Apprehension Scale (WAS). *[Database Record]. APA PsycTests.* https://doi.org/https://doi.org/10.1037/t73755-000

Deane, P. (2011). *Writing Assessment and Cognition*. Educational Testing Services, Research Report 11-14. http://www.ets.org/research/contact.html

Deane, P., Odendahl, N., Quinlan, T., Fowles, M., Welsh, C., & Bivens-Tatum, J. (2008). *Listening. Learning. Leading. Cognitive Models of Writing: Writing Proficiency as a Complex Integrated Skill*. http://www.ets.org/research/contact.html

Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, *32*(4), 521–541. https://doi.org/10.1177/0265532215575626

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*(2), 155–185. https://doi.org/10.1177/0265532207086780

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly*, *9*(3), 270–292. https://doi.org/10.1080/15434303.2011.649381

Ekens, T., & Meestringa, T. (2013). *Beoordeling van en Feedback op Schrijfvaardigheid. Een handreiking voor de tweede fase voortgezet onderwijs [Assessing and Giving Feedback on Writing. A tool for Upper Secondary Education]*. https://www.slo.nl/publicaties/@4216/beoordeling-feedback/

Elander, J., Harrington, K., Norton, L., Robinson, H., & Reddy, P. (2006). Complex skills and academic writing: A review of evidence about the types of learning required to meet core assessment criteria. *Assessment and Evaluation in Higher Education 31*(1), 71–90. https://doi.org/10.1080/02602930500262379

Elf, N., & Troelsen, S. (2021). Between joyride and high-stakes examination: Writing development in Denmark. In *International Perspectives on Writing Curricula and Development: A Cross-Case Comparison* (pp. 169–191). Taylor and Francis Inc. https://doi.org/10.4324/9781003051404-9

Elliott, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (2018). Accessible Instruction and Testing Today. In S. N. Elliot, R. J. Kettler, P. A. Beddow, & A. Kurz (Eds.), *Handbook of Accessible Instruction and Testing Practices Issues, Innovations, and Applications* (2nd ed., pp. 1–17). Springer International. https://doi.org/https://doi.org/10.1007/978-3-319-71126-3

Elosua, P. (2022). Validity evidences for scoring procedures of a writing assessment task. A case study on consistency, reliability, unidimensionality and prediction accuracy. *Assessing Writing*, *54*, 100669. https://doi.org/10.1016/J.ASW.2022.100669

Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, *32*(4), 365–387. https://blogs.baruch.cuny.edu/baruchteachingpracticum2015/files/2015/08/A-Cognitive-Process-Theory-of-Writing.pdf

Galbraith, D., & Rijlaarsdam, G. (1999). Effective strategies for the teaching and learning of writing. *Learning and Instruction, 9*(2), 93-108. https://doi.org/10.1016/S0959-4752(98)00039-5

Gamaroff, R. (2000). Rater reliability in language assessment: the bug of all bears. *System*, *28*(1), 31–53. https://doi.org/10.1016/S0346-251X(99)00059-7

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, *26*(4), 507–531. https://doi.org/10.1177/0265532209340188

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing*, *15*(2), 100–117. https://doi.org/10.1016/j.asw.2010.05.002

Graham, S. (2019). Changing How Writing Is Taught. *Review of Research in Education*, *43*(1), 277–303. https://doi.org/10.3102/0091732X18821125

Graham, S., Berninger, V., & Fan, W. (2007). The structural relationship between writing attitude and writing achievement in first and third grade students. *Contemporary Educational Psychology*, *32*(3), 516–536. https://doi.org/10.1016/j.cedpsych.2007.01.002

Graham, S., Gillespie, A., & McKeown, D. (2013). Writing: Importance, development, and instruction. *Reading and Writing*, *26*(1), 1–15. https://doi.org/10.1007/s11145-012-9395-2

Graham, S., & Perin, D. (2007). WritingNext: Effective Strategies to Improve Writing of Adolescents in Middel and High Schools. *A Report to Carnegie Corporation of New York*. www.carnegie.org/literacy.

Greenberg, K. L. (1992). Validity and Reliability Issues in the Direct Assessment of Writing. *Journal of the Council of Writing Program Adminstrators*, *16*(1–2), 7–23. https://associationdatabase.co/archives/16n1-2/16n1-2all.pdf

Hamp-Lyons, L. (2012). Writing teachers as assessors of writing. In Kroll, B., (Ed), *Exploring the Dynamics of Second Language Writing* (pp. 162–190). Cambridge University Press. https://doi.org/10.1017/cbo9781139524810.012

Hayes, J. R., & Flowers, L. S. (1983). *A Cognitive Model of the Writing Process in Adults. Final Report.* https://eric.ed.gov/?id=ED240608

Heidari, N., Ghanbari, N., & Abbasi, A. (2022). Raters' perceptions of rating scales criteria and its effect on the process and outcome of their rating. *Language Testing in Asia*, *12*(1). https://doi.org/10.1186/s40468-022-00168-3

Hendrix, T., & van der Westen, W. (2018). *Visie op het vak taal/Nederlands voor Curriculum.nu [Vision on the subject language/ Dutch for Curriculum.nu]*.

https://levendetalen.nl/wp-content/uploads/2018/01/Visie-op-het-vak-taal-Nederlands-voor-Curriculum.nu-Hendrix-en-Van-der-Westen-jan-2018-DEF.pdf

Heuvelmans, A. P. J. M., & Sanders, P. F. (1993). Beoordelaarsovereenstemming [Rater Agreement]. In T. H. J. M. Eggen & P. F. Sanders (Eds.), *Psychometrie in de Praktijk* (1st ed., pp. 443-469). Cito. https://ris.utwente.nl/ws/portalfiles/portal/282804368/KM1994046008.pdf

Homayounzadeh, M., Saadat, M., & Ahmadi, A. (2019). Investigating the effect of source characteristics on task comparability in integrated writing tasks. *Assessing Writing*, *41*, 25–46. https://doi.org/10.1016/j.asw.2019.05.003

Humphry, S. M., & Heldsinger, S. A. (2014). Common Structural Design Features of Rubrics May Represent a Threat to Validity. *Educational Researcher*, *43*(5), 253–263. https://doi.org/10.3102/0013189X14542154

Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, *60* (2), 237-263. https://doi.org/10.2307/1170611

Jia, W., & Zhang, P. (2023). Rater cognitive processes in integrated writing tasks: from the perspective of problem-solving. *Language Testing in Asia*, *13*(1). https://doi.org/10.1186/s40468-023-00265-x

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving Score Differences in the Rating of Writing Samples: Does Discussion Improve the Accuracy of Scores? *Language Assessment Quarterly*, *2*(2), 117–146. https://doi.org/10.1207/s15434311laq0202_2

Jönsson, A., Balan, A., & Hartell, E. (2021). Analytic or holistic? A study about how to increase the agreement in teachers' grading. *Assessment in Education: Principles, Policy and Practice*, *28*(3), 212–227. https://doi.org/10.1080/0969594X.2021.1884041

Kane, M. (2013). The Argument-Based Approach to Validation. *School Psychology Review* *42*(4), 448-457.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17. https://doi.org/10.1111/J.1745-3992.1999.TB00010.X

Kim, H. R., Bowles, M., Yan, X., & Chung, S. J. (2018). Examining the comparability between paper- and computer-based versions of an integrated writing placement test. *Assessing Writing*, *36*, 49–62. https://doi.org/10.1016/j.asw.2018.03.006

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing*, *35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing*, *18*(4), 300–308. https://doi.org/10.1016/j.asw.2013.09.003

Kraf, R., Lentz, L., & Pander Maat, H. (2011). Drie Nederlandse instrumenten voor het automatisch voorspellen van begrijpelijkheid - Een klein consumentenonderzoek [Three Dutch Tools for automatic prediction of understanding – A small consumer investigation]. *Tijdschrift Voor Taalbeheersing*, *33*(3), 249–265. https://kennisbank-begrijpelijketaal.nl/images/file/tools_tvt_def.pdf

Lederman, J. (2018). Writing Assessment Validity: Adapting Kane's Argument-Based Validation Approach to the Assessment of Writing in the Post-Process Era. *Journal of Writing Assessment*, *11*(1). https://escholarship.org/uc/item/1n22m978

Lee, Y. W., & Kantor, R. (2005). Dependability of new ESL Writing test score: evaluating priotype tasks and alternative rating schemes. *ETS Research Report Series*, *2005*(1), i–76. https://doi.org/10.1002/j.2333-8504.2005.tb01991.x

Livingston, S. A. (2018). Test Reliability-Basic Concepts. *EST Research Memorandum 18-01*. https://www.ets.org/Media/Research/pdf/RM-18-01.pdf

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing*, *30*(3), 329–344. https://doi.org/10.1177/0265532213480129

McCutchen, D. (1996). A Capacity Theory of Writing: Working Memory in Composition. In *Educational Psychology Review 8*(3), 299-325. https://doi.org/10.1007/BF01464076

Meadows, M., & Billington, L. (2010). *The effect of marker background and training on the quality of marking in GCSE English*. Centre for Education Research and Policy. https://api.semanticscholar.org/CorpusID:62899889

Meijerink, H. P., Letschert, J. F., Rijlaarsdam, G. C. W., van den Bergh, H. H., & van Streun, A. (2009). *Referentiekader taal en rekenen [Framework Language and Maths]*.

https://www.rijksoverheid.nl/onderwerpen/taal-en-rekenen/referentiekader-taal-en-rekenen

Moss, P. A. (1994). Validity in High Stakes Writing Assessment: Problems and Possibilities. In *Assessing Writing* 1(1),109-128. https://doi.org/10.1016/1075-2935(94)90007-8

Nederlands Nu!, & Sectie Bestuur Nederlands Levende Talen. (2018). *Advies examen Nederlands [Advice Examinations Dutch]*. https://lerarennederlands.nl/wp-content/uploads/2019/04/Advies-Examens-Nederlands-met-draagvlakonderzoek-finaal-1.pdf

Ono, M., Yamanishi, H., & Hijikata, Y. (2019). Holistic and Analytic Assessments of the TOEFL iBT® Integrated Writing Task. *JLTA Journal*, *22*(0), 65–88. https://doi.org/10.20622/jltajournal.22.0_65

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing*, *38*(4), 21–36. https://doi.org/10.1016/j.asw.2018.08.001

Palermo, C. (2022). Rater characteristics, response content, and scoring contexts: Decomposing the determinates of scoring accuracy. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.937097

Perelman, L. (2018). *Towards a New Naplan: Testing to the Teaching.* NSW Teachers Federation. https://www.nswtf.org.au/wp-content/uploads/2022/06/Towards-a-new-NAPLAN-Testing-to-the-teaching.pdf

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217–230. https://doi.org/10.1016/j.jslw.2013.02.003

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing*, *31*, 98–112. https://doi.org/10.1016/j.asw.2016.08.005

Rijlaarsdam, G., van den Bergh, H., & Couzijn, M. (2005). *Effective Learning and Teaching of Writing A Handbook of Writing in Education* (G. Rijlaarsdam, H. van den Bergh, & M. Couzijn, Eds.; 2nd ed.). Springer. https://doi.org/10.1007/978-1-4020-2739-0_1

Rooijackers, P. (2007). ‘We zetten in op schrijfvaardigheid’; schrijven waarschijnlijk terug in het eindexamen Nederlands [We go for writing; writing likely to return in central examinations of Dutch]. *Levende Talen Magazine*, *94*(5), 5–7. https://lt-tijdschriften.nl/ojs/index.php/ltm/article/view/316

Sabti, A. A., Md Rashid, S., Nimehchisalem, V., & Darmi, R. (2019). The Impact of Writing Anxiety, Writing Achievement Motivation, and Writing Self-Efficacy on Writing Performance: A Correlational Study of Iraqi Tertiary EFL Learners. *SAGE Open*, *9*(4). https://doi.org/10.1177/2158244019894289

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics, Vol. 1. Disorders of first-language development; Vol. 2. Reading, writing, and language learning* (Vol. 2, pp. 142–175). Cambridge University Press.

Schipolowski, S., & Böhme, K. (2016). Assessment of writing ability in secondary education: Comparison of analytic and holistic scoring systems for use in large-scale assessments. *L1 Educational Studies in Language and Literature*, *16*(1), 1–22. https://doi.org/10.17239/L1ESLL-2016.16.01.03

Schoonen, R. (1997). *Beoordeling van de samenvattingsopdracht in het nieuwe examen Nederlands [Assessing the Summary task in the New Central Examination of Dutch]*. Levende Talen Magazine, *84(516), 6-10.* https://lt-tijdschriften.nl/ojs/index.php/ltm/article/view/1082

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Schuurs, U. (2021, November). Het toetsen van schrijfvaardigheid. *Examens*. https://cito.nl/media/lp5pzz0g/cito_artikel_schrijfvaardigheid_in_examens.pdf

Skar, G. B., & Aasen, A. J. (2021). School writing in Norway: Fifteen years with writing as key competence. In J. Jeffery & J. M. Parr (Eds) *International Perspectives on Writing Curricula and Development: A Cross-Case Comparison* (1st ed., pp. 192–216). Routledge. https://doi.org/10.4324/9781003051404-10

Skar, G. B., & Jølle, L. J. (2017). Teachers as raters: An investigation of a long-term writing assessment program. *L1 Educational Studies in Language and Literature*, *17*(Specialissue). https://doi.org/10.17239/L1ESLL-2017.17.01.06

Slomp, D. H. (2012). Challenges in assessing the development of writing ability: Theories, constructs and methods. *Assessing Writing*, *17*(2), 81–91. https://doi.org/10.1016/j.asw.2012.02.001

Slomp, D. H., & Fuite, J. (2004). Following Phaedrus: Alternate choices in surmounting the reliability/validity dilemma. *Assessing Writing*, *9*(3), 190–207. https://doi.org/10.1016/j.asw.2004.10.001

Van Den Bergh, H., & Meuffels, B. (2000). Schrijfvaardigheden en schrijfprocessen [Writing Skills and Writing Processes]. In A. Braet (Ed.), *Taalbeheersing als communicatiewetenschap: Een overzicht van theorievorming, onderzoek en toepassingen.: Vol. Coutinho* (1st ed., pp. 122–153). https://www.researchgate.net/publication/46602244

Van der Leeuw, B., & Meestringa en Ravesloot, T. C. (2012). *Kijkwijzers SLO • nationaal expertisecentrum leerplanontwikkeling Beter zicht op het referentiekader taal*. Enschede: SLO. https://www.slo.nl/thema/vakspecifieke-thema/nederlands/referentiekader-taal/@4341/kijkwijzers-beter/

Webb, N. M., & Shavelson, R. J. (2005). Generalizability Theory: Overview. In *Encyclopedia of Statistics in Behavioral Science*. Wiley. https://doi.org/10.1002/0470013192.bsa703

Wiggins, G. (1994). The Constant Danger of Scarificing Validity to Reliability: Making Writing Assessment Serve Writers. *Assessing Writing*, *1*(1), 129–168.

Wind, S. A. (2019). Examining the Impacts of Rater Effects in Performance Assessments. *Applied Psychological Measurement*, *43*(2), 159–171. https://doi.org/10.1177/0146621618789391

Wind, S. A., & Engelhard, G. (2013). How invariant and accurate are domain ratings in writing assessment? *Assessing Writing*, *18*(4), 278–299. https://doi.org/10.1016/j.asw.2013.09.002

Zou, S. (2022). The Impact of Rating Scales on the CET-4 Writing: A Mixed Methods Study. In L. Hamp-Lyons & Y. Jin (Eds.), *Assessing the English Language Writing of Chinese Learners of English, pp 11 -28*. Springer. https://link.springer.com/book/10.1007/978-3-030-92762-2

10 Appendices

Appendix A

Rater Questionnaire

General Questions

1 How long have you worked as teacher of Dutch?
2 Which grades do you teach?
3 What do you think is important in writing lessons? Think about, for example, use of language, cohesion, referencing of sources, language conventions based on genre, knowledge of textgenres.
4 How do you translate your ideas about writing to your writing lessons?

Rating Procedures

1 How much time did you spend on rating?
2 Could you describe your rating process?
3 Could you make clear how you rated? For example, did you read all performances first and then start rating? Did you rate per sub domain?
4 Did you reconsider previously assigned scores for certain students? If so, why?
5 Did you miss essential aspects of writing on the rating form? Which and why?
6 Were there any sub domains that you would remove? Which and why?

Appendix B

Student Self-Perception Questionnaire

	This questionnaire is about the lessons prior to your Writing Exam, the lesson materials that were used during the lessons and the exam itself.	
	Item	Answer possiblities
1	What is your student code?	Open question
2	Prior to the school exam you have received a documentation file. How difficult were the sources in the documentation file?	1- very easy to very hard -5
3	To what extent were the sources in the documentation file useful for the exam?	1- very useful - completely useless - 5
4	How much time in hours did you spend on familiarising yourself with the sources in the documentation file?	open - limited to max 20 hours
5	How many additional sources about the subject have you looked for?	open - limited to max 10 sources
6	How would you describe your knowledge about the topic at the start of the lesson series?	very good - severely insufficient - 7
7	How would you describe your knowledge about the topic at the end of the lesson series?	1- greatly increased - remained the same - 4
8	What did you think of the topic?	1- very interesting - very boring - 5
9	How much have you learned from the following elements in the lessons about writing?	0- learned nothing - leaned a lot -5
9_1	naming text genres and purposes of sources in the documentation file	0- learned nothing - leaned a lot -5
9_2	scoring a ChatGPT expository essay	0- learned nothing - leaned a lot -5
9_3	making a skeletal outline for an expository essay	0- learned nothing - leaned a lot -5
9_4	receiving teacher feedback	0- learned nothing - leaned a lot -5
9_5	scoring the expository essay with the rubrics	0- learned nothing - leaned a lot -5
9_6	rewriting my draft version after having received feedback	0- learned nothing - leaned a lot -5
9_7	receiving classroom instruction about writing expository essays	0- learned nothing - leaned a lot -5
23	How stressful were the following elements for you?	0- very stressful - not stressful at all - 5
23_1	naming text genres and purposes of sources in the documentation file	0- very stressful - not stressful at all - 5
23_2	scoring a ChatGPT expository essay	0- very stressful - not stressful at all - 5
23_3	making a skeletal outline for an expository essay	0- very stressful - not stressful at all - 5
23_4	receiving teacher feedback	0- very stressful - not stressful at all - 5
23_5	scoring the expository essay with the rubrics	0- very stressful - not stressful at all - 5
23_6	rewriting my draft version after having received feedback	0- very stressful - not stressful at all - 5

	This questionnaire is about the lessons prior to your Writing Exam, the lesson materials that were used during the lessons and the exam itself.	
23_7	receiving classroom instruction about writing expository essays	0- very stressful - not stressful at all - 5
10	How easy would you say it is to give feedback on someone else's writing?	1- very easy - very hard -5
12	How useful was the peer feedback you received on your writing?	1- very useful - completely useless - 5
13	You have written a skeletal outline during the lessons. How did you experience creating the skeletal outline?	1- very easy - very hard -5
14	You have written a mock essay during the lesson series. How did you experience writing the mock essay?	1- very easy - very hard -5
24	Was your skeletal outline helpful during the exam?	1- a lot - not at all - 5
17	These questions are about the exam itself. Did you have enough time to complete the assessment?	yes - no
18	Why did you not have enough time?	open
19	How did you feel during the exam? To what extent did you experience one or more of the following problems during the exam:	
19_1	I had a black out	1- very relaxed - very stressed - 5
19_2	I did not have enough time to write down my ideas.	1- very relaxed - very stressed - 5
19_3	I am always stressed during exams.	1- very relaxed - very stressed - 5
19_4	I had not prepared well.	1- very relaxed - very stressed - 5
19_5	I did not understand the instructions in the task description.	1- very relaxed - very stressed - 5
19_6	I did not have any ideas at that moment.	1- very relaxed - very stressed - 5
25	To what extent do you agree with the following statements about the setting of the exam?	
	the rooms and space surrounding the room were quiet	1 - not at all - totally - 4
	there was an opportunity to ask questions	1 - not at all - totally - 4
	the digital environment worked well	1 - not at all - totally - 4
	it was clear what materials I was allowed to use during the exam	1 - not at all - totally - 4
21	What grade do you expect to receive for this exam?	open

Appendix C

The Questions as Presented to Students in the Writing Anxiety Test adapted from Daly & Miller (1975)

Writing Anxiety Test

(1) I avoid writing. (+)
(2) I have no fear of my writing's being evaluated. (-)
(3) I look forward to writing down my ideas. (-)
(4) I am afraid of writing essays when I know they will be evaluated. (+)
~~(5) Taking a composition course is a very frightening experience.(+)*~~
(6) Handing in a composition makes me feel good. (-)
(7) My mind seems to go blank when I start to work on my composition. (+)
(8) Expressing ideas through writing seems to be a waste of time. (+)
~~(9) I would enjoy submitting my writing to magazines for evaluation and publication. (-)*~~
(10) I like to write down my ideas. (-)
(11) I feel confident in my ability to express my ideas clearly in writing. (-)
(12) I like to have my friends read what I have written. (-)
(13) I'm nervous about writing. (+)
(14) People seem to enjoy what I write. (-)
(15) I enjoy writing. (-)
(16) I never seem to be able to write down my ideas clearly. (+)
(17) Writing is a lot of fun.(-)
(18) I expect to do poorly in composition classes even before I enter them. (+)
(19) I like seeing my thoughts on paper. (-)
(20) Discussing my writing with others is enjoyable. (-)
(21) I have a terrible time organizing my ideas in a composition course. (+)
(22) When I hand in a composition, I know I'm going to do poorly. (+)
(23) It's easy for me to write good compositions. (-)
(24) I don't think I write as well as most other people. (+)
(25) I don't like my compositions to be evaluated. (+)
(26) I'm not good at writing. (+)

1 = strongly agree
2 = agree
3 =uncertain
4 = disagree
5 = strongly disagree

* Questions removed from the original questionnaire

Appendix D

Original SLO Writing Rating Rubric for Dutch Expository Writing (Ekens & Meestringa, 2013)***.***

Rating Rubric for L1 writing assessment

	Scores
A. Coherence This is about the full text, the organization of the text in relation to the goal of the text, paragraph structure and relations between paragraphs. Think of: • Have thinking processes been described appropriately? • Does the thinking process contain enough steps? Are there sufficient, relevant steps in argumentation? Does the argumentation have a logic conclusion and are irrelevant steps in argumentation avoided? • Are the arguments connected appropriately?	1 2 3 4 5
B. Subject This is about whether the subject of text (the case or event) is dealt with in detail. Think of: • How well does the writer grasp the subject of his text? Howe well do they explain the subject of their text? Do they construct tension? Do they describe features in their argumentation? • To what extent does the writer engage the reader in their stories? Do they convince readers with their arguments? Do they objectively inform in factual texts?	1 2 3 4 5
C. Appropriate for genre and audience This is about lexical and grammar proficiency, relationships at sentences level and style in relation to context of the text, the genre, and the audience. Think of: • Have thinking processes been described appropriately? • Informal / formal, slang/ formal, daily / academic use? • Does the writer use appropriate wording for people's feelings and opinions, evaluations and words that strengthen or weaken? • Are the relations between inter- and intra-syntactic level sufficiently, logically presented?	1 2 3 4 5
D. Presentation This is about layout and language Think of: • Does the text show grammatical accuracy? Is there appropriate variation in sentences and phrases? • Does the text manifest accurate spelling? • How accurate and appropriate is interpunction? • Has the text been divided into paragraphs? Is the handwriting readable? Is lay out clear? How appropriate are illustrations and diagrams?	1 2 3 4 5

1 = fail; 2 = insufficient; 3 = at target level*; 4 = good; 5 = excellent.

* at target level for target level, for example 2F for 4 vmbo, above 2F for 3 havo./vwo, 3F for 5 havo; 4F for 6 vwo.