Validating integrated writing tasks –

A mixed-method approach to investigate the construct of summarization

Dissertation
zur Erlangung der Doktorwürde
durch den Promotionsausschuss Dr. phil.
der Universität Bremen

vorgelegt von Sonja Zimmermann

Erstgutachterin:
Prof. Dr. Claudia Harsch,
Universität Bremen

Zweitgutachterin:
Prof. Dr. Lia Plakans
University of Iowa

Datum des Promotionskolloquiums: 29.09.2022

ABSTRACT

Academic writing typically requires students to process information from different sources and integrate this information in their own texts. Hence, language tests for university admission purposes make increasing use of integrated writing tasks – tasks that provide students with language-rich source material (Knoch & Sitajalabhorn, 2013). Yet, the underlying construct of this kind of tasks is still an open issue, especially when looking at integrated writing tasks from an evaluation perspective: It is unclear which factors account for the performance, i.e. to what extent writing ability or reading skills contribute to the test results (Cumming, 2013; Weir, 2005).

The present study reports on validating an integrated writing task in the context of the new digital version of the *Test of German as a foreign language* (Test Deutsch als Fremdsprache; TestDaF) that became operational in late 2020. This task requires test takers to summarize information from a written text and a graphical input in relation to a given question.

Following recent approaches, the present study builds on a mixed-methods design to shed light on the construct underlying the integrated writing task by looking into a) the cognitive processes involved, b) the quality of the written products, and c) how reliable the integrated performances were scored.

In Strand 1 of this study, the cognitive processes of 19 international study applicants were examined, using a combination of eye-tracking and stimulated recall techniques. Findings from the quantitative and qualitative data revealed, that test takers engaged in a variety of cognitive processes related to basic processes of reading and writing, but also employed processes that integrated reading and writing in so-called shared processes. The identified

processes confirmed findings from previous research and allow for linking them to existing L2 integrated writing.

The written performances of the 19 participants from Strand 1 were also used in Strand 2 of the present study to investigate the relevance and accuracy of information included from the two sources, and to look into the transformation of the language of the input material. Findings not only showed differences between participants at distinct levels of proficiency in relation to source use and integration style, but also revealed differences in reproducing information between the written source text and the graphical input.

Strand 3 addressed the scoring of written performances. Applying many-facet Rasch measurement (MFRM), data of 445 examinees and 28 raters were analyzed. Results showed that raters were able to apply the rating scale as intended and rated consistently. To tackle the issue of the construct underlying the integrated writing task, and to measure the weight of reading and writing in integrated writing, a correlation analysis was used. The integrated writing task correlated highly with both the independent writing and the reading test, implying that both skills are involved in integrated writing assessment.

Following argument-based approaches to validation (e.g. Kane, 2013), the comprehensive empirical data gathered in all three strands contributed to establishing a validity argument for the integrated writing task of the TestDaF.

ON A PERSONAL NOTE

The last five years have been a rollercoaster ride. What started off with a pretty straightforward master plan quickly turned into a pretty "normal" PhD journey with many ups and downs – but: there is no "normal" in doing a PhD, especially not when you are a full-time working mom during a pandemic. Now that this journey has come to an end, I would like to thank family, friends and colleagues without their support and help I would have not been able to finish this endeavor.

A big shout-out to my supervisor Claudia Harsch who was willing to accompany my journey from the very beginning. Claudia, thank you for always believing in me, for your advice and support, for asking the right questions at the right time, for two fruitful writing retreats on Norderney, and for finding the right balance between letting me do things at my own pace and setting deadlines. You always stressed the importance of balancing academia and personal life, and I know that this has been a challenge for both of us during COVID. Your moral support was invaluable during the last 20 months.

I also want to express my gratitude to Lia Plakans who agreed to review this thesis and to be part of the doctoral committee. Lia, your work has been the foundation of my research, and I cannot thank you enough for your time, your feedback and your supporting e-mails. Whenever I was "down" on that rollercoaster ride, your kind words lifted me up.

This work would have not been possible without the backing of my colleagues at TestDaF-Institut. I especially want to thank Hans-Joachim Althaus, Gabriele Kecker and Thomas Eckes for giving me the opportunity to go on this journey. Achim – thank you for making it possible to take time off to write on the PhD. Gabi – thanks for the possibility to attend the EALTA 2009 conference. It was an eye-

opener for me! Thomas – thanks for reviewing countless conference and paper proposals and for your patience with me when it comes to statistics. Daniela and Leska – thanks for double coding the qualitative data. Frank – thanks for letting me figuring out the difference between t-tests and non-parametric equivalents on my own. Kai und Malte – thanks for many unforgettable lunches. To my other colleagues: Thanks for bearing with me and my obsession with integrated writing!

I also want to thank my sparring partners in the doctoral research colloquium at the University of Bremen. I will miss the Tuesday morning exchange with you. Anika and Voula – our online writing sessions (which sometimes were more like a therapy session) helped me to focus. You are incredible! Anika – this PhD would have not been on integrated writing without you. Thanks for pushing me in the right direction!

I had the opportunity to share findings of my research at numerous national and international conferences, and especially my colleagues in EALTA and ILTA have been witness of this journey from the beginning. With their valuable feedback they helped to shape this research study. Many of them also made memories: Jay and Spiros – your workshop on qualitative methods at EALTA 2009 in Turku was the starting point of my academic career! Lynda and Sara – you told me back in 2012 that I know almost everything about assessing writing. I am still not convinced that this is true, but thanks for the compliment! Benjamin, Kathrin, Franz, Thomas – thank you for a very special evening in Valencia! Slobodanka and Jamie – I cannot imagine better partners for getting lost in Paris!

My longest and dearest friend Catrin Albers and her family always welcomed me to their home when I was in Bremen. Catrin and Holger – thanks for several late nights, many bottles of wine and having Merit move out of her room when I stayed. You are family to me!

Last but not least, I want to thank my family. As a first-generation student, I thank my parents for the opportunity to go to university. Mama, Papa – thanks for giving me the freedom to figure out what I want to do with my life. And as a parent myself, I hope that I am an inspiration for my daughter in pursuing one's dreams and working hard to succeed. My husband was a single parent for countless weekends that I spent at my desk. Rena, Helge – sorry I did not always get my priorities straight in the last five years. This is for you!

CONTENTS

TABLES & FIGURES

List of Tables

List of Figures

INTRODUCTION

Context of this study

International students have to prove a certain level of language proficiency in German to enter institutions of higher education (HE) in Germany by taking one of the officially recognized language admission tests (see *Rahmenordnung über Deutsche Sprachprüfungen für das Studium an deutschen Hochschulen (RO-DT)*)[1]. The scores that test takers yield in these language tests are the basis for decisions by the admission bodies whether or not students can take up their studies. In this sense, language tests for admission purposes are high-stakes tests, i.e. the decisions based on the test scores have significant consequences for the individual. Tests that are used in the context of language admission to HE in Germany should therefore adhere to professional guidelines to assure a high quality of the test by providing validity evidence since validity is regarded as one of the most pivotal aspects of test quality (see e.g. the *Standards for Educational and Psychological Testing*, AERA, APA & NCME, 2014; for the concept of validity also see Chapelle, 2012; Eckes, 2015b; Kecker, 2010).

Following argument-based approaches in language testing (Bachman, 2005; Kane, 2013a), testing bodies are expected to support their claim that the test scores allow for inferences about the ability of the test takers in the target language use (TLU) domain. In the context of language admission to HE in Germany this would require test providers to demonstrate the following: A successful performance in the test allows test users like admission bodies to infer from the test score that the examinee has sufficient language

[1] https://www.hrk.de/fileadmin/redaktion/hrk/02-Dokumente/02-07-Internationales/Rahmenordnung_ueber_Deutsche_Sprachpruefungen_fuer_das_Studium_an_deutschen_Hochschulen__RO-DT__2020.pdf

ability to cope with communicative language tasks he/she encounters in real-life situations at university. The construct of a language test used within a special context therefore should as closely as possible reflect the language requirements of the TLU. Only then are the decisions based on the inferences derived from test scores valid.

One of the officially recognized tests for language admission to HE in Germany is the *Test Deutsch als Fremdsprache* (TestDaF), a standardized proficiency test that is administered worldwide in accredited test centers. Since the test was introduced in 2001 the paper-based version of the TestDaF has established itself as "the 'go-to'-assessment" (Norris & Drackert, 2018). With more than 446.000 candidates so far, it is "the German language test for university admission with the most participants" (g.a.s.t., 2020, p. 2).

The paper-based version of the TestDaF assesses language competence separately for four skills – reading, listening, writing and speaking. But language use at university is by no means separated in different language skills, on the contrary, it requires the integration of different skills since students need to process information from various sources (oral, written and/or visual) before they produce their own written or spoken texts. For example, students read literature, listen to lectures, take notes, and then write a term paper. Or they listen to a discussion in a seminar, and then respond to it.

This integrated language use was also confirmed by a needs analysis that was conducted by the TestDaF-Institut between 2010 and 2011. As part of the ongoing validation process for a high-stakes language exam, the aim of that study was to see whether the test, and especially the test format, still reflected and assessed the requirements of language use in academia. More than 120 university lecturers and over 1.300 international students in their first year(s)

of study took part in the needs analysis (Arras, 2012; Marks, 2015). On the one hand, results showed that the paper-based TestDaF covered relevant aspects of academic language use. On the other hand, it became apparent that the test tasks did not fully mirror TLU since complex language activities that require the processing of written and/or oral input were missing.

Therefore, the newly developed digital version of the TestDaF was developed and became operational in late 2020. The digital TestDaF not only includes independent tasks, it also comprises integrated task types to assess academic language competence in a more authentic way.

The current study focuses on the integrated writing task of the digital TestDaF that requires examinees to summarize relevant information from a written and graphical input with regard to a given question. As a validation study, it aims at providing evidence for the proposed score interpretations for the newly developed test version in the context of the intended test use, i.e. university admission.

Integrated tasks are broadly defined as „test tasks that combine two or more language skills to simulate authentic language-use situations" (Plakans, 2013, p. 1). With regard to writing, integrated tasks are defined as tasks in which candidates have to process „language-rich" input material and then have to use the information from written, aural and/or visual sources within their own writing by transforming the language of the input material (Knoch & Sitajalabhorn, 2013, p. 306). They are often used within the context of tests for academic purposes, particularly because of their authenticity (Cumming, 2013). Since these tasks are „text-responsible" (Leki & Carson, 1997, p. 41), writers have to develop an understanding of the input material and engage in content-responsible writing activities that are also required in academic writing. By providing source material, integrated writing tasks

reduce the impact of background knowledge which has often been criticized as a major disadvantage of more independent writing tasks like the *timed impromptu essay* (Weigle, 2002). They also seem to have a positive washback on preparatory language classes (Weigle, 2004), and „produce opportunities for language learning" (Cumming, 2014).

Despite these promises, and even though integrated tasks are increasingly used in language tests to address validity and authenticity issues, there are also some challenges related to their use in writing assessments. One of the major concerns about integrated writing tasks is related to the combination of different skills. The fact that such tasks measure not only the ability to write, but also require the processing of source material, raises the question to what extent the results yielded from integrated writing tasks depend on the ability to comprehend input material or the ability to write. For this reason, researchers have referred to the confounding measurement of integrated writing tasks as „muddied measurement" (Weir, 2005) or as „task dependencies" (Cumming, 2013). The request for a clear construct definition – which is the prerequisite for test validation – remains yet unresolved. Nevertheless, research has approached this issue from different perspectives, trying to shed light on the construct underlying integrated writing tasks.

Research perspectives in integrated writing assessment

So far studies have looked at reading-to-write tasks[2] from three different perspectives to investigate the underlying construct and

[2] In this study, the terms integrated writing, reading-to-write, reading-into-writing, or source-based writing are used synonymously.

provide evidence for validity claims of this task type: (1) a process-oriented perspective, (2) a product-oriented perspective, and (3) a scoring perspective. Each approach and relevant literature will be discussed in subsequent chapters, but a short overview of the different approaches should allow for relating the current validation study within the wider research area on integrated writing assessment.

Research studies in language testing and assessment that have looked at integrated writing tasks from a *process-oriented perspective* examined the cognitive processes and strategies test takers employ when working on reading-into-writing tasks. To analyze and describe the involved processes, many studies have used existing writing models, but also looked at processes that are involved in *discourse synthesis* (Spivey & King, 1989). For example, Plakans (2009a, 2009b) used think-aloud protocols to look at cognitive processes of writers while working on an integrated writing task. She found out that writers used discourse synthesis processes like selecting the relevant information from the sources, connecting single information across different sources, and lastly organizing, i.e. structuring and transforming all the information in relation to the overall writing goal of the task. Yang and Plakans (2012) showed that the more frequent use of these processes had a positive impact on the writing performance: Texts of writers who employed more discourse synthesis processes were rated higher. In the age of digitalization, more and more language tests are computer- or internet-based, requiring test takers to type their written answers on the computer. This raises the question, to what extend the medium, and especially the familiarity with the keyboard layout impacts cognitive processing and writing activities during task completion. In his studies on the TOEFL iBT, Barkaoui (2014a, 2015) found only a small effect of typing skills and keyboard

familiarity on task completion for the integrated writing task in comparison to the independent task of the TOEFL.

From a *product perspective*, studies analyzed the written products and investigated factors that affect the quality of integrated writing performances such as task-based variables and test-taker characteristics. Task-based variables take into account the topic, type, length and linguistic complexity of the source material, as well as the required genre. Test-taker characteristics on the other hand include personal variables like writing experience, age, or level of education. Many product-oriented studies in integrated writing assessment research have focused on test-taker characteristics, especially on language ability, mostly comparing the written output of low- and high-level learners. The differences in the quality of the written products were thereby examined by looking at source use, i.e. to what extent the relevant information from the input material was included, and how the language of the source material was linguistically transformed and integrated in the written products (e.g. Keck, 2006; Plakans & Gebril, 2013; Shi, 2004; Weigle & Parker, 2012).

Looking at reading-to-write tasks from a *perspective of scoring*, the interpretation of test scores remains an open issue: Does this task type measure writing only, or is the underlying construct the sum of the skills involved, e.g. reading and writing? Or does the performance allow for making inferences to a special reading-writing-relation that can be interpreted „as a reciprocal interaction between literacy skills, in which the basic processes and strategies used for reading and writing are modified by an individual's goals and abilities, and also by external factors" (Asención Delaney, 2008, p. 141)? Studies have come to contradictory results: Some have shown that test scores from reading-to-write tasks have a higher correlation with writing than with reading ability (Asención Delaney,

2008; Cumming et al., 2005; Gebril, 2009, 2010). On the other hand, there is evidence that integrated writing tasks are suitable for assessing reading as well as writing competence (Shin & Ewert, 2015).

What is currently missing is a fundamental theory or model of writing from sources in the L2 (see Knoch & Sitajalabhorn, 2013). Yu (2013a) and Cumming (2013) therefore have asked for a comprehensive foundation and more systematic research agenda into the construct of integrated writing tasks. This agenda should take into account the great diversity of integrated writing tasks, e.g. by a clear and transparent definition of the required genres like summarization (Yu, 2013a). Also, there is a great need for the development of special rating criteria that reflect the content-related processing as well as the transformation of the language of source material (Chan, Inoue, & Taylor, 2015; Knoch & Sitajalabhorn, 2013; Yu, 2013a).

Overall research goal and relevance of the current study

The overall aim of the current study is to contribute to the validation of the recently introduced digital version of the TestDaF by looking specifically into the construct underlying the integrated writing task included in the writing component. Adapting different lines of research, the study wants to investigate the integrated writing task from three different angles as described in the section above: from a perspective of cognitive processing, from a product-oriented perspective and from the perspective of scoring. By looking at the cognitive processes test takers employ during task completion, analyzing the written outcomes and by finally linking these data with test scores, the study aims at establishing a validity

argument for score interpretations of the digital TestDaF integrated writing task within the context of university admission in Germany.

Based on frameworks for argument-based approaches to test validation (which will be further described in Section 1.4) the study wants to support the following assumptions:

- **Assumption 1: The integrated writing task of the digital TestDaF elicits cognitive processes that are typical for writing from sources within the context of academic writing.**
  In language testing research, cognitive processes of test takers have been investigated to provide evidence of the cognitive validity of a certain language assessment (e.g. Shaw & Weir, 2007; Weir, 2005). The underlying supposition is, that if the observed processes correspond with those processes that writers would use in the TLU domain, the task will be useful in that sense that it is authentic and interactive (Bachman & Palmer, 1996).
  The current study therefore investigates the cognitive processes involved in completing the integrated writing tasks of the digital TestDaF, and links these processes to existing reading and writing theories, with a special focus on the interaction of reading and writing during task completion.

- **Assumption 2: The (successful) processing and transformation of the input material is evident in the written product.**
  Even though an analysis of written performances does not allow for making claims about the cognitive processes during the actual writing process (Brinkschulte, 2012), looking at the texts of test takers should reveal to what extent the relevant information from the sources were included, and if writers were able to use their own words for reproducing these information. To write a successful summary in the digital TestDaF, test takers not only

need to develop an understanding of the task demands, but they also need to fully comprehend the source material. Hence, the processing of information during task completion impacts the written outcome. And since information on the cognitive processes of the test takers is also available, product data can be linked back to process data.

- **Assumption 3: The quality of the written summary is reflected in the score.**

The construct of a language test should also be reflected in the rating scales to rate the written performances. Integrated writing tasks require rating scales that differ from those used to rate independent writing performances (Chan, Inoue, & Taylor, 2015; Knoch & Sitajalabhorn, 2013). They should take into account the processing of information with respect to content, and the transformation of language from the input material. Differences between various levels of ability should be evident and reflected in the scale. For example, lower proficiency levels are characterized by leaving out relevant information and by more direct copying from sources. On the contrary, summaries at higher proficiency levels should include more relevant information and use more paraphrases. In other words, summaries of high quality are characterized by high semantic, but low linguistic closeness to the original sources.

If the integrated writing task of the digital TestDaF task mirrors relevant characteristics of academic writing (Assumptions 1 & 2), and if the test scores are based on the quality of the written products (Assumption 3), the current study eventually argues that the integrated writing task of the digital TestDaF is a valid and reliable measure for academic writing competence that is required by international students entering institution of HE in Germany.

In order to provide evidence for these assumptions, the thesis encompasses different strands, adopting the three research perspectives on integrated writing assessment mentioned in the section above:

- **Strand 1** uses eye-tracking and stimulated recalls to investigate the cognitive processes of test takers.
- **Strand 2** analyses the written performances in order to see how correctly the information from the source material has been reproduced and to what extent the sources have been linguistically transformed.
- **Strand 3** looks into the scoring of these performances, and how the integrated writing scores relate to other variables like reading or writing ability.

Each strand calls for different types of quantitative or qualitative data. Therefore, the thesis uses a mixed-method research design (further described in Section 1.4.), with each approach adding value to the overall research aim (Creswell, 2009; Creswell & Plano Clark, 2018; Dörnyei, 2007).

The current study does not only provide validity evidence for the integrated writing task of the digital TestDaF, by doing so, it also addresses some existing research gaps in the field of integrated writing assessment.

First of all, most existing studies investigated the underlying construct of integrated writing tasks in the context of English as a foreign language (EFL). Integrated writing tasks used in tests for languages other than English have not been in the focus of language testing research yet. Investigating the cognitive processes of writers of German as a foreign language, and exploring how they correspond to theoretical models used in studies for EFL, could contribute to a more comprehensive model of writing from sources in the L2.

Furthermore, existing studies mainly looked into tasks with reading input only, or explored integrated reading-listening-to write tasks as used in the TOEFL iBT. So far, however, integrated writing tasks which require test takers to summarize relevant information from written and graphical input have not been investigated.[3]

But most importantly, most published research has investigated integrated tasks from only one of the three above mentioned perspectives, i.e. exploring the cognitive processes, analyzing the quality of the written performances, or looking at the scoring of integrated tasks. Often either the cognitive processes or the quality of the written product were related to each other or to the performance outcome, i.e. the score, respectively. But no study so far has examined the construct underlying integrated writing with a comprehensive approach by linking all three perspectives.

Structure of the thesis

Chapter 1 provides the theoretical background of this study and addresses two areas: 1) the TestDaF integrated writing task, and 2) frameworks for test validation.

Regarding the task, the chapter reflects on the importance of summarization within academic writing and takes a critical look how this is operationalized in different language tests for admission purposes, including the digital TestDaF. Because the test format requires test takers to type their summary on the computer, the effect of the medium will also be explored. The chapter closes with the description of the applied framework for test validation, and with the presentation of the overall research design.

[3] Whether or not summary tasks with only visual input (like IELTS Academic, Writing Task 1) can be regarded as an integrated writing task can be questioned. See chapter 1.1 for a broader discussion on that issue.

The following chapters (chapters 2 – 4) independently deal with Strands 1 to 3 respectively to investigate the construct underlying the integrated writing task of the digital TestDaF. Each of the three strands is conceptualized as a research study in its own, therefore each chapter (a) provides a theoretical background, (b) reviews the relevant research literature and (c) derives emerging research questions before (d) explaining the research methodology and finally (e) describing and discussing the results for each approach taken, also (f) taking into account the limitations.

Chapter 2 covers the process-oriented Strand 1 to test validation, investigating the cognitive processes test takers employ when working on the summary writing task of the digital TestDaF. Using a combination of eye-tracking and stimulated recall, the analysis of viewing behavior and cued retrospective interviews allows for further insights into reading and writing activities involved in integrated writing tasks. The analysis of eye-tracking measures focuses mainly on the time participants spent in *areas of interest* (AOIs) and on the transitions between the AOIs at different stages of the writing process. Detailed results from this quantitative analysis will be backed up by participants' quotes from the verbal reports, showing the engagement of test takers with the source material during task completion. The chapter also takes a closer look at the relationship between the cognitive processes and test taker characteristics like typing skills and level of language competence.

The characteristics of the written summaries are the main focus of Chapter 3. The detailed qualitative analysis of this strand draws on existing coding schemes for integrated writing performances and is also informed by the TestDaF rating scale to link the analysis to performance outcomes. Exemplified by test takers performances, this strand explores how writers of different performance levels

include relevant content-related information from the sources, and by what means they transformed the language of the input material.

Chapter 4 focuses on the analysis of rating integrated writing performances. To investigate the extent to which the TestDaF integrated writing scores are related with writing or reading ability, integrated writing performance data was related to test takers' scores in an independent writing task and results from a reading comprehension test. The use of *many-facet Rasch measurement* (MFRM) also reveals insights into the reliability of test scores, taking into account the ability of test takers, the difficulty of the task, as well as the leniency or harshness of individual raters.

Chapter 5 brings together results from all three perspectives by reviewing the claims stated in the introduction against the background of the findings from eye-tracking, stimulated recalls, text analysis, and scoring.

The final Chapter 6 presents implications for teaching and learning in the context of preparatory language classes, as well as implications for rating scale design and rater training.

Validation is simple in principle,
but difficult in practice.
(Kane, 2011, p. 15)

1 VALIDATING THE INTEGRATED WRITING TASK OF THE DIGITAL TESTDAF

This chapter provides relevant background information for the overall research study, focusing on aspects that are related to the task as well as on the validation framework used in this study.

The chapter reflects on the importance of summary writing within the academic context and takes a closer look at how this specific kind of writing is operationalized in different language tests, specifically in the digital TestDaF. It will then present findings from existing research on score and construct equivalence of computer- and paper-based language tests to explore possible effects of the medium. Another focus of this chapter is the validation framework applied in this study. Together with the task related aspects, it forms the basis for the overall research design which will be presented at the end of this chapter.

1.1 Summarization as an integrated writing task

Writing is an essential part of academic studies at institutions of HE – not only in Germany. Students take notes and add personal comments to handouts or scripts while listening to lectures or participating in seminars. They use these notes to write protocols or to prepare themselves for written assignments or tests. In many fields of study, particularly within the humanities or social sciences, written term papers are still an obligatory part of the curriculum in Germany (Ehlich & Steets, 2003). To write these papers, students read extensive scientific literature, sometimes including statistical or visual sources, and excerpt or rather verbalize (in the case of non-

written sources) relevant information before they finally write their own texts by transforming the input material into their own words.

These examples demonstrate two things: First of all, there is a great variety of written texts students have to produce during their studies. The texts do not only vary in terms of expected genre and text length, but also according to formal requirements which are different for each subject. Furthermore, academic writing is not only text production, it also requires to a great extent receptive skills, i.e. the comprehension of textual and/or visual sources. Not only native speaker students struggle with these requirements of academic writing (see Dittmann, Geneuss, Nennstiel, & Quast, 2003), especially international students encounter difficulties with writing academic texts in their L2. For one thing, they have to make an enormous effort to read and comprehend complex academic literature, and secondly they have to search for appropriate expressions to reformulate the processed information in their own words to avoid the allegation of plagiarism which requires a broad range of linguistic resources many non-native speakers may lack (Grießhammer, 2011; Stezano Cotelo, 2003).

Nonetheless, a common feature of academic writing is the processing and transformation of knowledge, and summarizing being a core skill within that whole process (Grabe & Zhang, 2013; Hood, 2008). Thereby, important information is separated from less important information, details are left out, so that by shortening and condensation the source text will be transformed into a whole new text (Keseling, 1993; Hirvela & Du, 2013). The target text still has some similarities with the source text, but the processed information is not simply mechanically reproduced, but rather reformulated and restructured, transforming the source qualitatively. In this sense, summarizing is an essential language function within academic

writing, and summarization as a writing genre can be regarded, similar to the excerpt, as an *assisting genre*.[4]

Synthesis writing, a specific kind of summarizing, places the same reading-writing demands on the writer, but additionally involves the processing and linking of main ideas across sources (Grabe & Zhang, 2013; Hirvela, 2004).

Hence, writing tasks which require students to summarize or synthesize written texts and information from aural or visual sources are authentically mirroring the requirements of academic writing (Cumming, 2013). For this reason, these kind of writing tasks are commonly used in language tests for academic purposes, but operationalized in a variety of ways (Knoch & Sitajalabhorn, 2013; Yu, 2013b).

For example, in the integrated writing task of the *Test of English as a Foreign Language, Internet-Based Test* (TOEFL iBT), test takers have to read a text, and then listen to a lecture. In their written response, which should be between 150 and 225 words long, they should „summarize the points made in the lecture, being sure to explain how they cast doubt on specific points made in the reading passage."[5] The *Pearson Test of English Academic* (PTE Academic) operationalizes summarization differently. Test takers read a passage of approximately 300 words and have to write a one-sentence summary of no more than 75 words, including the main points of the reading text.[6] The Academic Writing test of the *International English Language Testing System* (IELTS) contains one task where test takers have to summarize information from two

[4] In his discourse on academic genres in the context of higher education in Germany, Ehlich (2003) uses the term „Hilfstextart" (2003, p. 23).
[5] https://www.ets.org/toefl/ibt/prepare/practice_sets/writing; retrieved 09.12.2019.
[6] https://pearsonpte.com/the-test/format/english-speaking-writing/summarize-written-text/; retrieved 09.12.2019.

graphs in at least 150 words. This includes describing relevant information, focusing on the main features and – if necessary – comparing information across sources.[7] In the writing component of the French *Diplôme Approfondi de Langue Française* (DALF) on the level C1 test takers are required to synthesize information from different text material. Therefore, they have to select relevant information by identifying a common theme in all the provided material, before presenting this information in their own writing which should be around 220 words long.[8]

It becomes evident that there is a great variety of these tasks that require some sort of summarization, and which are usually subsumed under the umbrella term of *integrated writing tasks*. The variety is not only caused by differences in type and length of input material (ranging from short to extended reading passages or graphical input), but also by discrepancies between the expected outcomes (single sentence summaries or texts up to 200 words). Looking at the variance with regard to input and output, some of these tasks even lack key premises to be defined as integrated writing tasks. Knoch and Sitajalabhorn (2013) call for a more focused definition of integrated writing tasks, demanding „the stimulus materials […] to provide sufficient language (either in written or audio format) to allow writers to produce sufficient text to be rated by assessors“ (Knoch & Sitajalabhorn, 2013, p. 304). Following this definition, tasks that only use visuals as input (like e.g. IELTS Academic Writing, task 1) cannot be considered as integrated since the input material does not include „a significant proportion of language” (ibid., p. 304).

[7] https://takeielts.britishcouncil.org/take-ielts/prepare/free-ielts-practice-tests/writing/academic/task-1; retrieved 09.12.2019.
[8] https://www.ciep.fr/sites/default/files/migration/delfdalf/documents/DALF_C1_exemple2.pdf; retrieved 09.12.2019.

Even though there are differences across task types, the cognitive processes involved in summarization as a „discourse in its own right“ (Yu, 2013b: 97), always include source text comprehension, as well as the reduction and reconstruction of the main ideas (Yu, 2013b).

1.2 The integrated writing task of the digital TestDaF

As described in the introduction of this thesis, results of a needs analysis (Arras, 2012; Marks, 2015) revealed the demand for revising the test format of the paper-based TestDaF. New task types, including integrated tasks, were developed to mirror complex language use at university more closely. The test development process applied international standards for quality assurance and comprised intensive trialing (Kecker, Zimmermann, & Eckes, in press). The test construct is based on a model of communicative language competence initially proposed by Bachman (1990), and more recently taken up in the *Common European Framework of References for Languages* (Council of Europe, 2001). It also takes into account typical communicative tasks that students encounter at institutions of HE in Germany.

The digital TestDaF consists of four components, i.e. a reading, a listening, a writing and a speaking section. Test taker's performance in each of the four components is related to one of three TestDaF levels (*TestDaF-Niveaus*, TDN) of language proficiency – TDN 3, 4 or 5. These level correspond to the CEFR levels B2 to C1.[9] Eligible for admission to institutions of HE in Germany are test

[9] The correspondence of the TestDaF-levels to the levels B2 and C1 of the CEFR was recently confirmed in a Standard Setting for the digital TestDaF (see https://www.testdaf.de/de/ueber-testdaf/arbeiten-mit-gast/aktuelles/neuigkeiten/; retrieved 09.10.2021).

takers who yield at least the TestDaF level 4 in every section of the test.

The test covers topics from different fields of subject like humanities and social sciences, natural sciences, engineering as well as economics and medical science. Due to the heterogeneous target population, it is not related to a specific curriculum or language course. Topics and texts need to be comprehensible for a non-specialized audience; the test tasks mirror relevant language skills that are essential across disciplines.

The test is delivered through a special exam security software (*Safe Exam Browser*, SEB[10]) which puts the computer in a kiosk mode. By this, certain functionalities of the computer are temporarily restricted, e.g. candidates cannot log on to the internet or use communication tools like text chats. This allows for monitoring that examinees do not use non-permitted resources during task completion.

In the writing component of the digital TestDaF test takers have to prove that they can master relevant writing functions that are part of different texts and genres required within academia. In order to do so, they have to produce coherent and well-structured texts, thereby determining the outline and organization of their own writing, and make revision where necessary. In their texts they should phrase their own ideas and viewpoints, make references to ideas and views of others, and summarize information from different sources.

The writing component consists of two task types – one independent and one integrated writing task.[11] The integrated writing task requires test takers to synthesize information from a

[10] https://safeexambrowser.org/about_overview_en.html.
[11] The independent task is used as an instrument for the scoring approach and will be described in more detail in the according chapter 4.3.3.

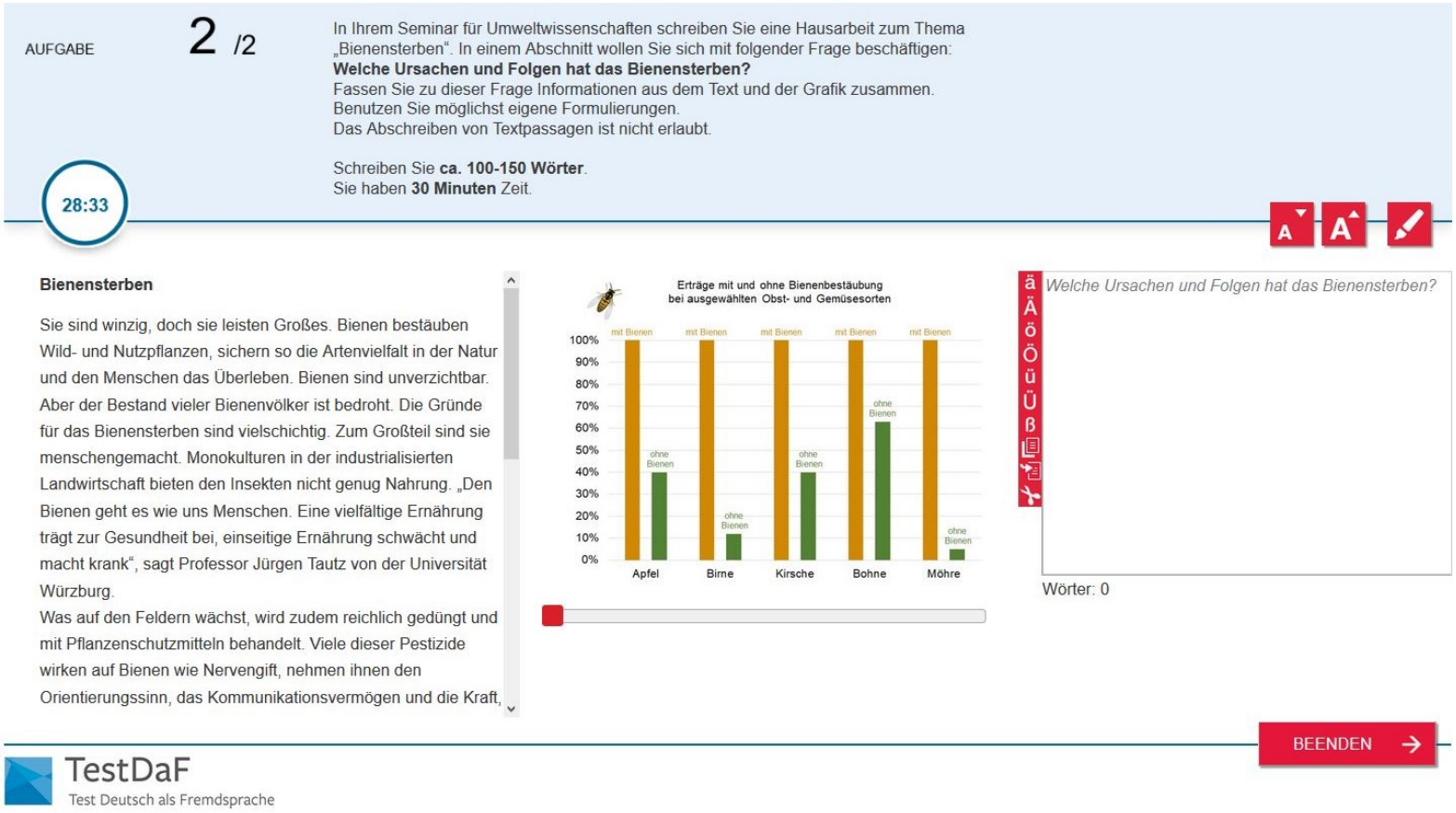
written text and a graphical input in relation to a given question, which is stated in the instructions, and also in the text box at the beginning of the task. As soon as participants click in the text box and start writing, the question disappears. The instructions also provide the situational embedding of the task: The expected summary is intended to be a section within a chapter of a written assignment at university[12]. The written input text is approx. 250-300 words long, and examinees have to scroll in order to read the whole text. The graphical input contains either supplementary, contradictory or redundant information. Test takers have to contrast and compare both sources, before synthesizing relevant information with regard to the given question, thereby reducing information from the input material. They are allowed to use key terms from the original sources for reproducing relevant ideas, but they are not allowed to lift longer passages.

The digital test environment allows for adjusting the font size, i.e. to enlarge the text, as well as to zoom into the graphical input. Test takers can also highlight text in the instructions and the reading text, but they cannot copy and paste from the input material since the SEB does not allow for this shortcut and blocks the right mouse klick.

Time for task completion is 30 minutes, test takers are expected to write between 100 and 150 words. They can control the number of words they have produced so far by a word count; a timer shows the remaining time. Figure 1-1 shows the graphical user interface for the integrated writing task.

[12] See Yu (2013b) for the need of specific task instructions for summary writing in testing contexts.

Figure 1-1 *Model task of the integrated writing task in the digital TestDaF*[13]



[13] This model task is publicly available on www.testdaf.de.

1.3 Effects of the test delivery mode: Writing online

With the advancing digitalization in the last decade, computer familiarity and typing skills have become essential in occupational and academic settings. Especially in academic contexts, students basically read and write texts on the computer, except maybe for personal note-taking, so that the construct of L2 academic writing should be further expanded and should consider keyboarding skills and computer literacy as an integral part (Barkaoui & Knouzi, 2018; Jin & Yan, 2017). Already in 2006, Chapelle and Douglas pointed out that with this development, the testing of L2 writing in traditional paper-based tests might even introduce a potential bias, and that computer-based testing hence allows for more authentic task design. More and more large-scale, standardized language assessments offer computer-based versions of their tests, but the equivalence of both test delivery modes has often been questioned.

Research in the field of writing assessment has therefore tried to demonstrate two kinds of equivalence of both testing modes, i.e. score and construct equivalence. Studies with a focus on the former mainly examined the effect of delivery mode on test-taker scores (e.g. Brunfaut, Harding, & Batty, 2018; Chan, Bax, & Weir, 2018; Jin & Yan, 2017; Weir, O'Sullivan, Yan, & Bax, 2007), while studies with a focus on the latter looked into the differences in cognitive processing in computer- and paper-based writing assessments. With the increasing use of integrated writing tasks, recent research has explored the effect of the writing medium for different task types, i.e. integrated and independent writing tasks (Brunfaut et al., 2018).

By investigating the score and construct equivalence, studies have also looked into variables like computer familiarity, typing skills or test takers' perceptions on the cognitive processes (e.g. Barkaoui, 2015), and their effect on the quality of the written

performances (Barkaoui & Knouzi, 2018), as well as the interaction between those factors and language proficiency (Barkaoui, 2014a). Finally, some research exists on the relationship between the type of keyboard test takers used during task completion and test scores (Ling, 2017a, 2017b), finding no significant effects, but revealing that test takers would prefer to use a familiar keyboard layout in a test-situation (Ling, 2017b).

A limitation of all the above described studies is that they used self-reported data on computer familiarity and typing skills which may differ from the actual ability. Barkaoui (2014a, 2015) and Barkaoui and Knouzi (2018) therefore used direct measures like a typing test to examine the effect of keyboarding skills on the writing process and the performances.

In his studies on the independent and integrated writing task of the TOEFL iBT, Barkaoui (2014a; 2015) examined the effects of delivery mode, language proficiency and typing skills on test takers cognitive processes and test scores. Regarding test scores, Barkaoui (2014a) found only a small effect of typing skills, dependent on the task type with scores on the independent task more affected by computer writing skills. On the contrary, language proficiency contributed substantially to variance in scores on both task types. In terms of cognitive processing, Barkaoui (2015) could provide evidence that test takers engaged in cognitive processes as expected from theoretical writing models. For example, in the stimulated recalls test takers reported that the integrated tasks involved more source-based activities, whereas the independent task required them to generate their own ideas, to do more planning, and to revise their texts more often.

Looking at the effect of delivery mode and the influencing variables on the quality of written products, Barkaoui (2016), Barkaoui and Knouzi (2018) and Jin and Yan (2017) reported that

test takers wrote longer texts and made less language errors on the computer, but results were mainly related to the overall language proficiency of the participants. While these studies focused on independent writing tasks, Kim, Bowles, Yan, and Chung (2018) compared the quality of a paper- and a computer-based integrated writing test. They also confirmed that computer-written essays were slightly longer, but otherwise could not observe substantial differences in the quality of the performances between the two test versions.

Overall, results on the effects of writing mode are not consistent. While most studies that looked into score equivalence reported on no significant differences between the two writing modes, some studies reported on small effect sizes favoring the paper-based test (e.g. Brunfaut et al., 2018), whereas some results showed significant higher scores of computer-based writing (Jin & Yan, 2017). Studies also looked into the impact of different variables on the scores, e.g. keyboarding skills, computer familiarity or test takers perception of computer-based tests. In general, these variables did not have any significant impact on the scores, but small differences in cognitive processing (Chan, Bax, & Weir, 2018) and linguistic features of the written products (Barkaoui & Knouzi, 2018) could be observed, so that construct equivalence of the two delivery modes should be questioned.

The current study does not look into score or construct equivalence of the paper-based and the digital TestDaF. But findings from previous research offered useful insights on writing online, and provided the basis for the research design of this thesis.

1.4 Argument-based approaches to validation in language testing

This dissertation builds on an argument-based approach to validation. One key element of argument-based validation frameworks like Kane's *Interpretation Use Argument* (*IUA*; Kane, 2011, 2013) or Bachman's *Assessment Use Argument* (*AUA*; Bachman, 2005) is the call for validating the interpretation and uses of test scores rather than only validating the test itself (e.g. Borsboom & Markus, 2013).[14]

In order to do so, Kane's argument-based framework consists of two steps: an *interpretive argument*, and a *validity argument*.

> An *interpretive argument* specifies the proposed interpretations and uses of assessment results by laying out a network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the assessment scores (Kane, 2011, p. 8).

A schematic overview of the interpretive argument and its central components is provided in Figure 1-2. The *target domain* provides the broader context to the interpretive argument. It defines the real-life domain in which the ability to be tested can be observed, and presents the background to the interpretation and uses of test scores.

The first inference (*scoring*[15]) in the interpretive argument is made by translating the *observation*, i.e. the test-taker's performance, into a score by means of rating and statistical procedures. Based on the assumption that the *observed score* is derived from a representative sample of tasks (i.e. the *universe of generalization*), the *generalization inference* allows for drawing

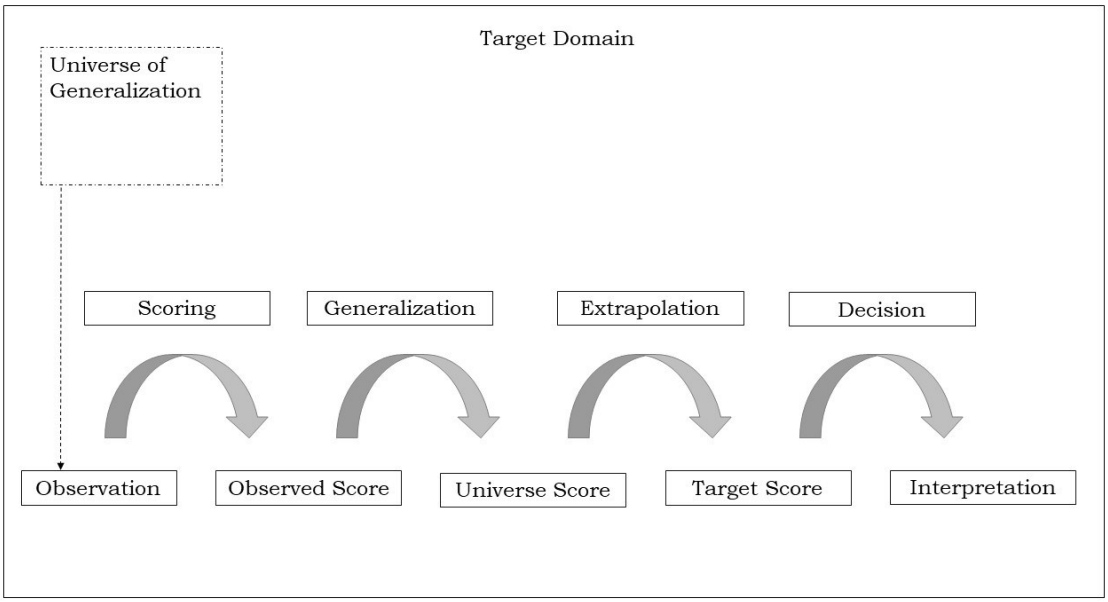
[14] The idea that not a test itself is valid, but rather the interpretations and uses of test scores, goes back to Cronbach (1971), but is mainly associated with Messick (1989). For a more comprehensive discussion on concepts of validity see also Chapelle (2012), Eckes (2015b), or Kecker (2010)

[15] The scoring inference is also referred to as 'evaluation inference' (see Knoch and Chapell, 2018).

conclusions about a test takers' performance in a larger domain of tasks. Through the *extrapolation* inference, assumptions are then made about a test-taker's ability in the target domain. The score is then interpreted by test users which finally leads to a decision, e.g. if a test taker is admitted to university or not.

According to Kane (2013b), many of the assumptions that the inferences rely on can be taken for granted, especially in a high-stakes context. The focus should therefore be on the ones that seem problematic.

Figure 1-2 *Interpretive argument*

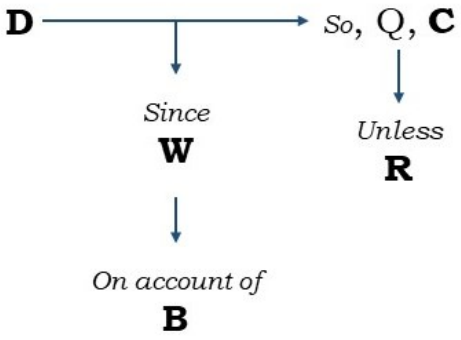


After the interpretive argument has laid out chain of reasoning of "what is being claimed" (Kane, 2011, p. 4), the *validity argument* "provides an evaluation of the interpretive argument's coherence and the plausibility of its inferences and assumptions" (ibid, p. 8).

For doing so, Kane and others (Bachman, 2005; Mislevy, Almond, & Lukas, 2003) draw on Toulmin's structure of arguments (Toulmin, 2003). Following Toulmin's logic (see Figure 1-3), every *Claim* (**C**), i.e. a decision to be made, is based on *Data* (**D**). The link

between the data and the claim is justified by a *Warrant* (**W**) and the evidence supporting this interpretation (*Backing*, **B**). The inference made from the data to the claim can be rebutted (**R**) by counterclaims, in some cases it might be necessary to qualify (Q) the degree to which the intended inference holds true.

Figure 1-3 *Toulmin's argument structure*



Throughout this dissertation, support for the different inferences will be collected to evaluate the credibility for the overall claim that scores derived from the integrated writing task of the digital TestDaF and the according rating instruments and procedures allow for making inferences about academic writing ability of test takers (Chapelle, Enright, & Jamieson, 2010; Chapelle & Voss, 2014).

The intended validity argument for the TestDaF integrated writing task would require backing for the following assumptions:

- Scoring inference: The evaluation of the integrated writing performances is based on rating criteria that capture relevant characteristics of the performances, as well as using sound and reliable scoring procedures.
- Generalization inference: The observed scores can be generalized to a greater universe of expected scores, i.e. across task versions and across raters.

- Extrapolation inference: The observed scores can be extended to the target domain, i.e. the construct underlying the integrated writing task is related to writing ability in the TLU domain.
- Decision inference: The test scores can be interpreted meaningfully by the test users, and are used appropriately.

While the thesis focuses on the first three inferences, i.e. scoring, generalization and extrapolation, providing support for the decision inference is not in the scope of this dissertation, mainly because the responsibility for validating the interpretations of test score and uses is shared by the test provider and score users like policy makers and admission offices at university (Deygers, 2017).

1.5 Overall research design

Mixed-method research designs have gained popularity in applied linguistics and language assessment research recently (Moeller, Creswell, & Saville, 2016). Studies using this specific research approach combine elements of quantitative and qualitative methods to overcome the weaknesses of a single approach. Thus, mixed-methods are more than simply collecting and analyzing quantitative and qualitative data. Using both kind of data allows researchers to investigate a complex issue from multiple perspectives, hence increasing the validity of research outcome by triangulation (Dörnyei, 2007).

The current study applied a *convergent* (Creswell & Plano Clark, 2018)[16] or *concurrent* (Dörnyei, 2007) design consisting of three

[16] In their latest edition, Creswell and Plano Clark showed how their typology of the convergent design changed over the years from "concurrent triangulation strategy" (Creswell, 2009) over "convergent parallel design" (Creswell and Plano Clark (2011) to "convergent design" (Creswell and Plano Clark, 2018), with a clear focus now on the intent, and not on the timing of the data collection and analysis.

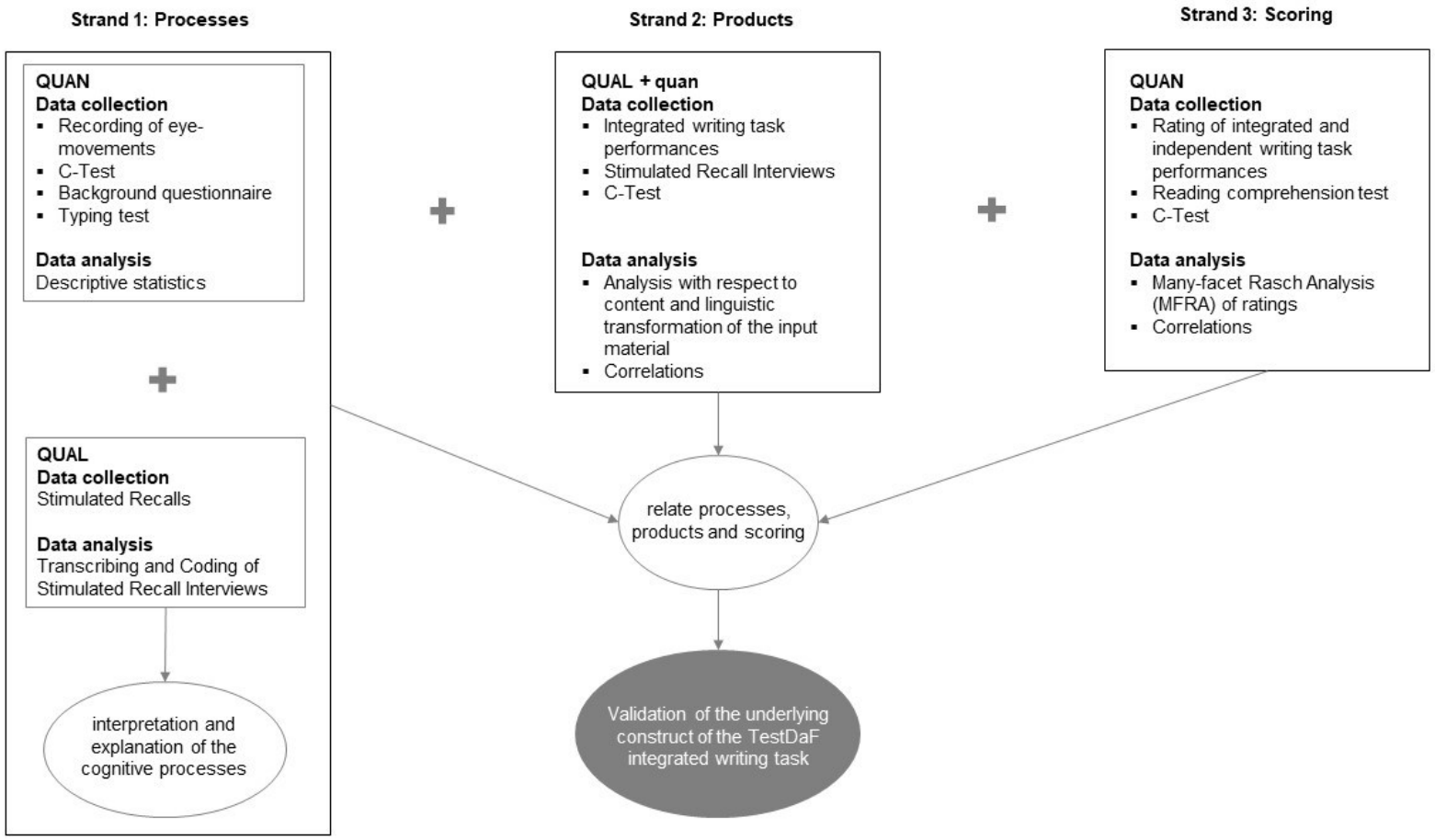
separate strands, each using a different approach to investigate construct underlying the integrated writing task of the digital TestDaF (see Figure 1-4). All data was collected simultaneously, and each strand was analyzed separately. Results were integrated at a later interpretation and explanation phase.

Strand 1: The first strand made use of quantitative (QUAN) eye-tracking and qualitative (QUAL) stimulated recalls to investigate what cognitive processes test takers engage in when writing from sources.

Strand 2: A qualitative (QUAL) analysis of the written texts was used to explore how the sources were transformed linguistically and with respect to content. An additional correlation analysis (quan) should inform about the linking of product and process data, i.e. if the quality of the written output was related to viewing behavior of participants.

Strand 3: Quantitative data of the performance ratings and other variables were used to examine a) the reliability of ratings, and b) to look into possible effects of factors like reading comprehension or overall language proficiency on the scoring of integrated writing performances.

Figure 1-4 *Overall research design*



1.5.1 Participants

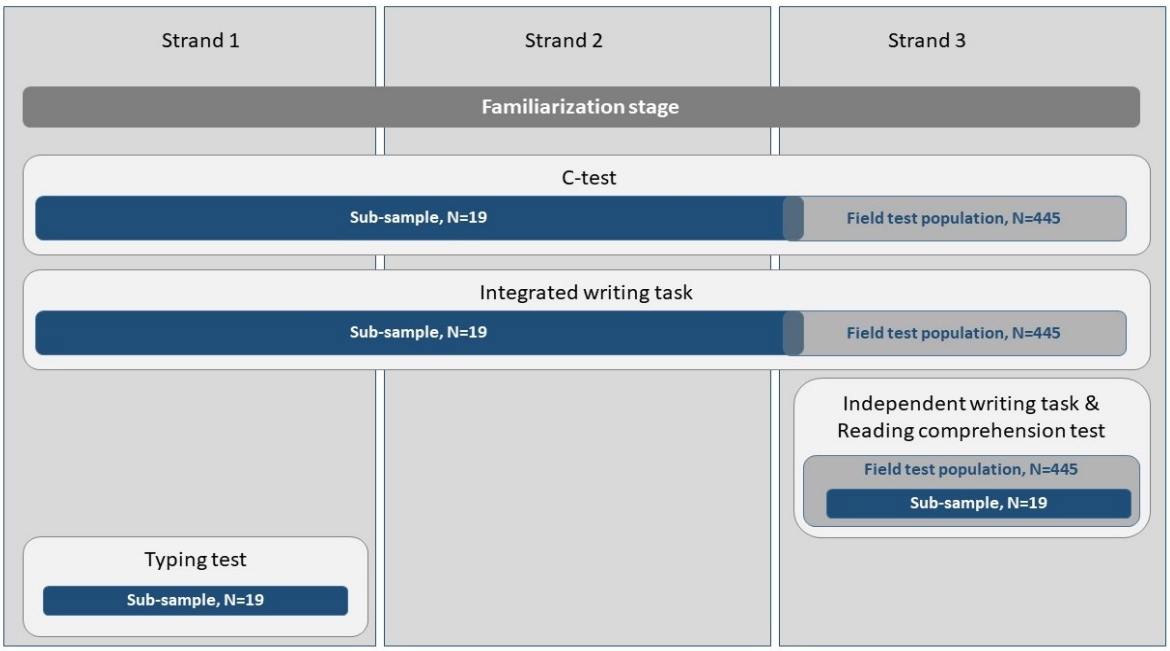
Data for the study was collected during a piloting stage for the digital TestDaF in June 2018. A representative sample of 445 participants in test centers across the world took part in this piloting stage. The newly developed test tasks were piloted in two different test versions (Set 1 and Set 2). Data from field test participants were used to look into the scoring of the integrated writing task (see Strand 3; Chapter 4). A detailed description of this sample including the distribution of the participants across Set 1 and Set 2 can be found in section 4.3.2.

To relate the cognitive processes of examinees during task completion and their written outcomes to the scoring of their performances, the design of this study used a nested sample within the larger sample of the whole field test population (see Creswell & Plano Clark, 2018). This sub-sample used in Strand 1 and Strand 2 consisted of 19 participants (for a detailed description see section 2.3.4) and allowed for an in-depth analysis of the cognitive processes and the written performances.

1.5.2 Instruments and data collection

Besides the integrated writing task, the mixed-method design called for different instruments that were used within in each strand (see Figure 1-5 for an overview): (1) a C-test, (2) a typing test, (3) a reading comprehension test and (4) an independent writing task.

Figure 1-5 *Overview of instruments*



Since the C-test is the only instrument that was used across strands, it will be shortly described in this section, while a detailed description of the other instruments will be given in the according chapters of the different strands.

(1) C-test results were used in all three strands of this study to relate the cognitive processes, the integrated writing performances of participants as well as the scoring of their writing to their overall language proficiency. To place candidates on the global scale of the *Common European Framework of Reference for Languages (CEFR)*, four texts from the calibrated item bank of the onSET German[17], an online language placement test, were given to all participants. Unlike the original onSET (Eckes, 2010), the C-test version used during field testing included only four instead of eight texts, consisting of 20 gaps each. All

[17] More information and a sample test can be found on http://www.onset.de.

participants worked on the same four texts in the same order. Each of the four texts was assigned to one of the levels A2-C1 of the CEFR, and texts were presented with increasing difficulty. Participants had a maximum of 5 minutes to complete each text.

(2) In Strand 1, participants also completed a typing test to investigate whether their keyboarding skills affected their (cognitive) processing while working on the integrated writing task of the digital TestDaF (as described in section 1.1).

(3) For Strand 3, a reading test was used to measure the test takers' reading ability in order to look at the effect of reading competence on integrated writing performances.

(4) Participants also worked on the independent writing task of the digital TestDaF to compare independent and integrated writing performance in Strand 3.

Both, the reading comprehension test and the independent writing task will be further described in section 4.3.3.

The actual data collection for all three strands was preceded by a familiarization stage, in which all participants had access to a practice test of the reading and writing component of the digital TestDaF to familiarize themselves with the test tasks. This included short videos for each task, in which the graphical user interface (GUI) was explained and candidates were given instructions on what they had to do in order to complete the task. They could also access model tasks and work on them, but no feedback on their results was provided. Due to practical constraints, the familiarization process was not controlled, and no data was collected at this stage.

Data for Strand 3 was collected during field testing in licensed test centers worldwide (see section 4.3 for more details). Data for the sub-sample used in Strand 1 and Strand 2 was collected using

convenient sampling of participants who were enrolled in language courses at a languages center at a large German university. Even though only the C-test and the typing test were of special interest for these two strands, the data collection for this smaller sample also intended that participants worked on the reading comprehension test and the independent writing task. This allowed for this data to be included in the data analysis of Strand 3 and for linking the data of all three strands. Details of the data collection for this smaller sub-sample are described in section 2.3.6.

1.5.3 Data analysis

The different types of data used in each strand call for different types of analysis which will be addressed in the respective chapters. All quantitative data – eye-tracking data, C-test results and text scores – were analyzed using SPSS (version 24.0.0.1) and FACETS (version 3.80.1). The stimulated recall interviews were transcribed, and coded along with the written performances using NVivo 12. Parts of the qualitative data were double coded and inter-coder reliability was checked to assure the reliability of the coding schemes used.

Writing is best understood as a set of distinctive thinking processes which writers orchestrate or organize during the act of composing.
(Flower & Hayes, 1981, p. 366)

2 A PROCESS-ORIENTED APPROACH TO VALIDATION[18]

The following chapter focuses on the cognitive processes that test takers employ when working on the integrated writing tasks of the digital TestDaF. By doing so, it aims at providing evidence for the cognitive validity of the test task. The underlying assumption is that the integrated writing task of the digital TestDaF elicits cognitive processes that correspond with "a postulated theoretical construct" (Yang, 2014, p. 314) of the TLU ability.

The theoretical foundation that the current study builds on will be described in the introductory section of this chapter, followed by a review of relevant literature that has investigated cognitive processes in integrated writing assessment. Section 2.3 describes the methodology applied in the current study; findings are reported in section 2.4. The chapter closes with a discussion of results, including limitations of the current study.

2.1 Reading-writing processes in integrated writing assessment

Although a growing body of research has investigated the processes involved in integrated writing assessment (see section 2.2 for a more detailed literature review), a comprehensive model for writing from sources in the L2 is still lacking (Hirvela, 2005; Knoch & Sitajalabhorn, 2013).

[18] Some parts of this chapter have already been published in Zimmermann (2020a, 2020b).

Integrated writing certainly involves processes of reading and writing, but research has shown that the construct of reading-to write tasks is not simply the sum of reading and writing constructs (Wolfersberger, 2013). It is rather unique and can best be "conceptualized as a reciprocal interaction between literacy skills, in which the basic processes and strategies used for reading and writing are modified by an individual's goals and abilities, and also by external factors" (Asención Delaney, 2008, p. 141).

Often cited cognitive models of writing like the one by Flower and Hayes (1980) have some limitations in capturing the processes involved in source-based writing since they do not elaborate on the process of source-reading. Hayes stressed how vital critical reading skills are for the writing process in his '*New framework for understanding cognition and affect in writing*' (Hayes, 1996). In his updated model, Hayes perceived reading as a central part of revision, but he also stressed the importance of two other kinds of reading in the writing process: *reading source texts* and *reading to define tasks*. The latter is important for the writers to align their writing with the task requirements, i.e. to understand the question and what the expected genre is. According to Hayes, writers often fail to fulfill the task because of misunderstanding certain terms in the instructions like „interpret" or „argue" (Hayes, 1996: 20). In his model, Hayes also included the composing medium to the task environment, noting the effects the medium can have on the writing process, in particular on planning and revising (Hayes, 1996).

Even more integrated models like Kintsch's and van Dijks *Model of text comprehension and construction* (1978) have some shortcomings. In their model, Kintsch and van Dijk describe three operations that are involved in text comprehension and construction:

In a first step, the reader builds a mental representation of the text, before secondly condensing it to its gist. In a final third step, new texts are generated from what has been comprehended before. The model though is limited in the way that it was built on recalls and summarization protocols of proficient L1 writers. Kintsch's and van Dijk's participants did not have access to the text they had to summarize during writing – something that was key for their experiment since they were interested in text reproduction from memory. However, this is very different from typical source-based writing in university contexts where writers usually have access to the source material throughout the whole writing process.

Originated in reading research, and applied in the context of L1 writers, the model of *discourse synthesis* (Spivey & King, 1989) has been a valuable framework for investigating cognitive processes in integrated writing tasks. Discourse synthesis is a meaning making process in which readers become writers by synthesizing information from multiple sources (mainly written texts) in order to create new texts (Spivey & King, 1989; Spivey, 1990).

Reading and writing in discourse synthesis are not consecutive processes, where readers/writers first read other texts before starting to write their own text, it is rather a more hybrid act of literacy in which both processes co-occur:

> When writers compose from sources, reading and writing processes blend, making it difficult, if not impossible, to distinguish what is being done for purposes of reading from what is being done for purposes of writing. Although we see evidence of organizing, selecting, and connecting, we often cannot say whether a writer performs a certain operation to make meaning *of* the text that is read or to make meaning *for* the text that is being written. (Spivey, 1990, p. 258)

Three processes are involved in discourse synthesis: *selecting*, *organizing* and *connecting*. When reading for understanding, readers *select* content from a text – based on the demands of the task. They

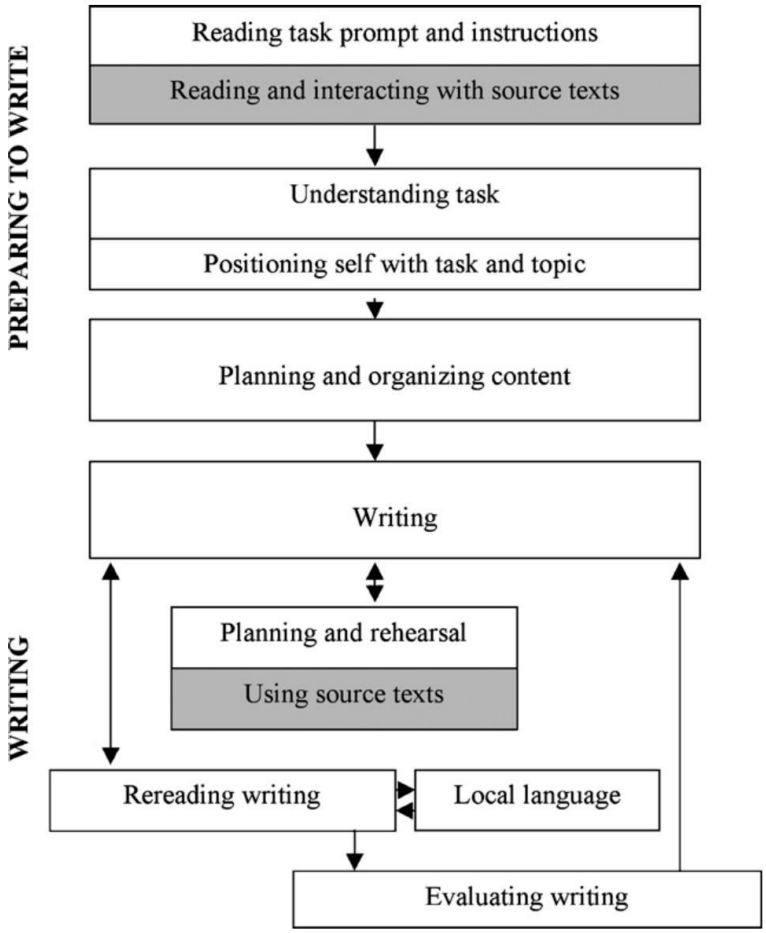
organize the content of the sources by applying organizational patterns, and they *connect* the information in the text on a local level. When readers become writers in source-based writing, they also apply these three operations. In source-based writing, writers organize their texts by applying a new organizational structure to the content of the sources, i.e. they organize by restructuring it – except in isomorphic summary writing, where writers often reproduce the organizational pattern of the source, simply "putting the content in a more compact but nevertheless similarly shaped package" (Spivey, 1990, p. 265). For writing their own text, they select relevant content from the sources, comparing if information is only mentioned once, or is repeated across sources. The selection of information is either driven by "textual relevance", i.e. the importance of the information in the text, or "contextual relevance" (van Dijk, 1979; cited from Spivey, 1990), i.e. the importance of information in relation to accomplishing a certain task. Finally, writers connect related ideas across sources, sometimes making inferences to previously acquired knowledge. Discourse synthesis can be seen as an act of comprehending *and* composing, in which the reader/writer builds a mental representation of sources and transforms them to create a new text by applying composing processes, including planning and revising.

Building on the discourse synthesis framework, Plakans (2008) developed a model for composing process in reading-to-write tasks (see Figure 2-1) on the basis of think-aloud protocols from L2 writers.

The composing process in Plakans's model consists of two distinct stages. While preparing to write, writers read the prompt and the instructions as well as the input material in order to understand the task and position themselves. They also engage in planning processes like organizing their text. The subsequent stage

of writing is characterized by a combination of writing, planning, interacting with the sources and an evaluation of the text written so far. While Plakans described the pre-writing stage as a more linear process, the writing stage itself is more recursive.

Figure 2-1 *Plakans's model for composing process for reading-to-write tasks*



Plakans argues that a writer in integrated writing assessments needs to employ the processes of discourse synthesis, i.e. organizing, selecting and connecting, and that her model allows for capturing these processes.

The discourse synthesis model has been used as a theoretical framework for research that has investigated cognitive processes involved in integrated writing assessment. This also becomes evident

in the next section which will provide an overview of the existing body of research in this area

2.2 Investigating cognitive processes in integrated writing tasks

Whereas back in 1985 Kennedy could claim that research "pays little attention to the processes of 'writing from sources'" (ibid., p. 437) and that "writing about reading sources is a fertile untilled area for future research" (ibid., p. 453), the interest in cognitive processes in source-based writing has been growing ever since.

Previous research has explored processes in integrated writing assessment from different perspectives. One strand of research has looked into cognitive processes involved in integrated writing in contrast to independent writing. In her study on a paper-based English placement test at a university, Plakans (2008) used think-aloud protocols as well as pre- and post-test interviews to investigate cognitive processes involved in integrated and independent writing. Results showed that processes in both task types differed. While the independent writing task elicited a more linear approach of planning, translating and revising, processes in the integrated writing task were not that straightforward. Writers spent less time planning before writing on the integrated task, but they seemed to do more planning during writing. She concluded that the reading-to-write tasks elicits more composing processes that can be linked to the discourse synthesis model (see section 2.1). In relation to the integrated writing task, the study also revealed that experienced and interested writers were more engaged with the source text compared to other writers.

Barkaoui (2015) used stimulated recalls to examine writing activities in the TOEFL independent and integrated writing tasks,

and to investigate how these activities changed over time during task completion. He also looked into the effects of language proficiency and keyboarding skill, as well as the relationship between writing activities and text quality, i.e. test score. In relation to writing activities by task type, results showed that both writing tasks elicited a wide range of writing activities like planning, evaluating and revising, whereby for the integrated task participants reported most frequently on interacting with sources (like referring to the source to retrieve main ideas). In relation to writing activities across time, Barkaoui found that participants predominantly adopted a rather linear approach for writing with the different task types – regardless of level of language proficiency and keyboarding skills. For the integrated writing task, this meant participants tended to read and interact with sources at the beginning of the writing process, before starting to plan, generate and actually write. Evaluation and revision were mainly taking place at the end of the writing process. While the relationship between writing activities and task scores was not significant, participants with higher scores were interacting more frequently with the writing task, especially interacting with the sources at the beginning of the writing process.

Other studies have looked into the role that reading and discourse synthesis strategies played in reading-into-writing tasks. For example, Plakans (2009b) looked into reading-related strategies of international students working on an integrated writing task for an EFL placement test. Think-aloud protocols as well as pre- and post-test interviews were analyzed for strategies that occurred during different stages of the writing process, i.e. pre-writing, writing, and revision. The analysis revealed that writers most commonly used comprehension strategies. Reading-related strategies were used during the whole writing process, but most strategies were employed during pre-writing. Results also showed

differences across writers, with higher-scoring writers using more reading strategies than lower-level writers, focusing on more global strategies for the purpose of the task. Low-scoring writers instead employed more word- and sentence-level strategies.

Plakans (2009a) used think-aloud protocols to investigate discourse synthesis processes in an integrated writing assessment. The think-aloud protocols (TAPs) were transcribed and coded for relevant discourse synthesis processes, i.e. *organizing*, *selecting* and *connecting*, as well as for other categories like monitoring and language difficulties. Results revealed a variation in composing processes across writers, with only four of the six writers engaging in discourse synthesis processes. Interview data and TOEFL scores revealed that these differences might be related to L2 proficiency and/or writing across academic fields. Those participants with lower TOEFL scores and less experience in synthesizing sources for their academic course work interacted less with the sources and employed fewer discourse synthesis processes.

To investigate the shared processes of reading and writing and to shed light on the integration of these two skills in integrated writing, Plakans and her colleagues (Plakans, Liao, & Wang, 2019) developed an iterative integrated writing assessment. TAPs of L2 university students showed that writers engaged in processes of reading and writing, while the task also elicited test-taking strategies and shared processes. Findings also revealed two shared key processes in reading and writing: comprehension and a focus on words, the latter confirming previous findings on 'bottom-up' processing in integrated writing assessment (e.g. Plakans, 2009).

Plakans and Gebril (2012) examined how writers were using the sources in argumentative integrated writing tasks. Through questionnaire, think-aloud and interview data they could show that writers engaged with the sources during the whole writing process.

Writers used the source texts for the purpose of "*academic writing and thinking processes*" (ibid., p. 29), i.e. they generated ideas by reading the sources and returned to the sources during writing to provide evidence for their opinions. Participants also read the sources "*to support language in writing*" by looking for language and terminology they could use.

While the above mentioned studies used think-aloud data, Li (2014) employed a post-test questionnaire on reading and writing strategies to look into strategy use during a summary task, which was delivered to 60 EFL learners from China. However, the questionnaire was also based on TAPs and retrospective interviews that had been conducted in a first step. Results revealed participants used both reading and writing strategies during task completion. Confirming Plakans's findings (Plakans, 2009b), Li reported that the most used reading strategies were related to bottom-up processes to comprehend the source text, while the most frequently used writing strategy showed that participants relied heavily on the source text. Looking at the relationship between reading and writing strategies, participants' summarization performances and their language proficiency, statistical analyses showed that writing strategies contributed more to the summarization performance than reading strategies, and that the performance was only weakly related to English proficiency.

Also working with self-reported data, but with a larger sample size, Yang and Plakans (2012) used structural equation modeling (SEM) to investigate strategy use of L2 writers on an integrated reading-listening-writing task and how it is related to test performance. Based on a strategy inventory that participants had to fill in after completing the integrated writing task, Yang and Plakans identified three factors that explained the complex nature of integrated writing strategy use: discourse synthesis strategy use

(DSS), self-regulatory strategy use (SELFS), and strategy use that is related to test-wiseness (TWS). Results showed that self-regulating strategies monitored other strategy use, and that DSS positively affected test takers' overall L2 writing ability and reduced direct copying from the sources. On the other hand, TWS had a negative impact on the integrated writing performances. Yang and Plakans concluded that "to perform well, test takers have to actively interact with the source texts by selecting, organizing, and connecting information, rather than relying on copying, patchwriting, or applying templates to writing" (Yang & Plakans, 2012, p. 93). Similar findings were reported by Yang (2014) who also used a strategy inventory and SEM to look into the relationship between strategy use und performance on a summarization task of Taiwanese EFL learners.

While the aforementioned studies relied on self-reported data only, recent studies apply a mixed-method design, combining eye-tracking metrics with self-reported data, to look into the cognitive processes. Wang (2018) looked into test takers processes in the context of a business English proficiency test in China to provide evidence for the cognitive validity of the test. He recorded the eye movements of participants while they completed an integrated reading-to-write task, and used these recordings as a stimulus for retrospective interviews. In addition, Wang used a process questionnaire (adopted from Chan, 2013) to collect responses from a larger sample of participants. Results confirmed that participants employed a variety of cognitive processes that could be linked to existing writing models (Hayes, 1996; Shaw & Weir, 2007) and the model of discourse synthesis. According to the stimulated recalls, participants most frequently used *selecting* and *micro-planning* as processes. The eye-tracking metrics revealed that the engagement with the input material during task completion differed between the

various sources: While some sources received high amounts of attention, participants looked at others less frequently. The varying degree of time participants spent looking at the different sources was also related to participant's reading behavior, i.e. reading carefully for comprehension before writing, and on the other hand reading expeditiously when searching for lexical support or specific content information during writing.

Michel et al. (2020) examined cognitive writing processes in the independent and integrated TOEFL iBT task, focusing on pausing behavior at different stages of the writing process. In their study, Michel and colleagues used data triangulation of keystroke logging, eye-tracking and stimulated recalls. Partially in line with previous research (Barkaoui, 2015; Plakans, 2008), their findings revealed that apart from source text use in the integrated writing task, participants' overall writing behavior was similar in both task types. Looking at cognitive processes at different stages of the writing process, Michel and colleagues could show that for the integrated task, participants were engaging in reading the source text at the beginning of task duration. In the middle stages text constructing processes became more important, involving higher- and lower order writing processes, whereas at the last stage participants focused on revision and monitoring processes.

Even though not situated in a testing context, Wolfersberger's (2013) qualitative study in the context of a classroom-based reading-into-writing assignment provided valuable insights into task representation. Task representation helps writers to "create an understanding of what skills, products, and processes the task requires and make a plan of action that will lead to a written product that appropriately fulfills the writing task" (Wolfersberger, 2013, p. 52). By investigating the writing process of four international EFL students with semi-structured interviews and classroom

observations as well as by analyzing the written drafts, Wolfersberger showed that task representation is an important step in the writing process that is created in two stages. In stage 1, participants drew on their existing writing experience "to facilitate an initial conceptualization of the task" (Wolfersberger, 2013, p. 60), while in the second stage task representation was shaped by the current writing context. Influencing factors like teacher feedback and classroom activities cannot be applied to standardized testing, but Wolfersberger's results also revealed that the source texts also played an important role in shaping the task representation – something that is also relevant for testing contexts.

In summary, existing research on L2 writing processes in integrated writing assessment has revealed the following: Cognitive processes involved in writing from sources differ from those involved when participants work on independent writing tasks. This is mainly related to interacting with the source material. Writers employ reading and discourse synthesis strategies when writing from sources throughout the whole writing process. Taking a closer look at different stages of the writing process, participants engage more with the source material before starting to write, whereas towards the end of the writing process, they are more engaging in revision processes, mainly focusing on their own writing. Studies also indicated that differences in cognitive processing involved in integrated writing are related to L2 proficiency. Higher proficient, i.e. higher-scoring students interact more frequently with the source material and employ more discourse synthesis strategies.

Since the integrated writing task of the digital TestDaF also contains graphical information, the next section briefly looks into the processing of visual sources in comparison to text-based sources.

2.2.1 Processing of graphical information

Studies on the use of graphs in L2 language assessment mostly exist in the context of speaking and listening assessment (Ginther, 2001; Katz, Xi, Kim, & Cheng, 2004; Xi 2005, 2010). Research on the processing of visual information in integrated writing research is relatively scarce (Yang, 2016; Yu, He, & Isaacs, 2017; Yu, Rea-Dickins, & Kiely, 2007).

The theoretical foundation of this existing research is the knowledge-based construction-integration model by Freedman and Shah (2002). According to this model, three dimensions affect graph comprehension and interpretation: (1) domain knowledge, i.e. the necessary content knowledge to correctly interpret the displayed characteristics, (2) graphical literacy skills, and (3) explanatory skills, i.e. the ability to verbalize and explain visual data.

In their study on the IELTS Academic Writing Task 1, Yu and colleagues (Yu et al., 2007) used TAPs, post-task interviews and questionnaires on graph familiarity to investigate how the use of different graphs affected cognitive processes of test takers during writing, and to what extent graph familiarity, the test takers' writing ability and a short training affect their cognitive processes during writing. It was found that participants' performances varied significantly depending on the type of graph that was used. It seemed that line graphs were easier to understand than statistical tables, participants received the highest mean scores when working on tasks including a line graph, while the mean scores for tasks including statistical tables were the lowest. Candidates also perceived line graphs to be easier, since they were the most familiar type of graph to them. While graph familiarity had no effect on task performance, participants expressed some psychological impact of graph familiarity on their processing. Yu and colleagues could also

show that a short training had a strong influence on the performances of candidates. In a follow-up study using a combination of eye-tracking, stimulated recalls and focus group discussions, Yu, He, and Isaacs (2017) basically confirmed the findings from the previous study.

The reported studies in the context of the IELTS have one limitation: the tasks participants worked on were graph-based only. Research looking into in the processing of visual information in combination with textual information and the differences between the two types of sources only exists in the context of reading research (Schnotz et al., 2017; Zhao, Schnotz, Wagner, & Gaschler, 2014). These studies used eye-tracking to explore whether the usage of text and the usage of pictures differ when secondary school students in Germany worked on a reading task to acquire knowledge. Results showed that there is a substantial difference in text processing and picture processing, depending on contextual factors of when the question was presented, i.e. before or after students had access to the text-and-picture material. Texts were used to construct meaning, whereas pictures were used for specific task-oriented purposes.

Even though Plakans (2009) had called for research that requires different modalities and/or different genres like summarization, most research still focuses on reading-to-write tasks that involve argumentation. Another shortcoming of the existing body of research in this area is that most studies have been conducted in EFL contexts, calling into question the transferability of insights on integrated writing processes to other foreign language settings. And finally, most existing process-oriented studies on integrated writing tasks have looked into the processing of written and/or spoken sources only, not taking into account visual sources.

2.3 Methodology

2.3.1 Research aims and questions

The main focus of this strand is to investigate the processes test takers engage with when completing the digital TestDaF integrated writing task. A mixed-method approach consisting of a combination of eye-tracking and stimulated recall was used to collect quantitative and qualitative data to answer the following research question (RQ):

RQ1: What are the cognitive processes that test takers use when working on the summary writing task of the digital TestDaF?

To be more specific and gain further insights into test takers' cognitive processing, the following sub-questions were formulated

RQ1a: How do test takers approach the task?

RQ 1b: How do test takers engage with the task, especially with the input material? Are the involved cognitive processes related to reading, writing, or are there specific 'integrated' processes?

RQ1c: Do the cognitive processes vary in relation to different stages of the writing process?

RQ1d: Are the cognitive processes affected by test-taker characteristics like language competence or typing skills?

RQ1e: Are the involved processes generalizable across different versions of the task?

2.3.2 Eye-tracking

Eye-Tracking has become quite popular in the field of applied linguistics in the past decade (Godfroid & Hui, 2020; Godfroid, Winke, & Conklin, 2020). In the context of language testing and assessment, studies made use of eye-tracking to look into the

cognitive processes of test takers while reading (Bax, 2013; Bax & Chan, 2016; Brunfaut & McCray, 2015; McCray & Brunfaut, 2018) or writing (Révész, Michel, & Lee, 2017; Yu et al., 2017).

For process-oriented studies, eye-tracking is a valuable tool since it provides a direct measure of cognitive processing of test takers while working on test tasks without interfering with the writing process (like e.g. TAPs). It is based on the assumption that eye movements are linked to cognition (Conklin, Pellicer-Sanchez, & Carroll, 2018; Holmqvist et al., 2011). According to this so-called *eye-mind-hypothesis*, or *eye-mind link*, it is „generally considered that when we measure a fixation, we also measure attention to that position“ (Holmqvist et al., 2011, pp. 21–22). Longer fixations usually mean more processing effort, while shorter fixations indicate less effort required to process the fixated item (Conklin & Pellicer-Sánchez, 2016).

Fixations, as well as *saccades* and *regressions*, are movements by the eye in response to a textual or written input that an eye-tracker can record (see Figure 2-2).

Figure 2-2 *Key eye-tracking measures*[19]

Eye-tracking is used in this study because it can inform about the viewing behavior of participants while working on the integrated writing task of the digital TestDaF, and hence can provide insights into the cognitive processing during task duration. Unlike verbal

[19] Figure taken from Brunfaut and McCray (2015).

protocols, eye-tracking is a concurrent method to record writers' mental activities without distracting participants from writing.

The eye-tracking measures used in this study are fixation-based measures and focus on the time participants spend in different AOIs[20] on the computer screen (see sections 2.3.6 for more details). The reason to focus on greater AOIs instead of single words is the difference between the digital test environment and experimental settings in reading research. Besides the fact that the text participants have to read is much longer (250-300 words) in comparison to the textual input participants have to process in reading research that uses eye-tracking. The GUI of the digital TestDaF also creates some challenges: The font size of the source text is smaller and the spacing between the lines is much more narrow than recommended for reading research. Participants also need to scroll to read the whole text. The present study does not focus on single words, the focus is to measure the viewing behavior in relation to larger AOIs, i.e. to measure the engagement of participants with the source material, e.g. if participants looked more frequently at some AOIs than others, and hence if they processed the single parts of the input material differently. Hence, the collected eye-tracking data still can provide valuable insights into the cognitive process in integrated writing.

A basic limitation of using eye-tracking technology is the fact that it can only provide evidence that someone looked at a certain area. And even though there is proof that fixation-based eye-movement measures are good indicators of cognitive processes, „it is impossible to tell from eye-tracking data alone what people think" (Holmqvist et al., 2011, pp. 71–73). It is therefore recommended to use method triangulation, e.g. by a combination of quantitative eye-

[20] Sometimes AOIs are also referred to as *region of interest* (ROI).

tracking measures and a qualitative description of thought processes through think-aloud or retrospective stimulated protocols (Godfroid, Winke, & Conklin, 2020).

2.3.3 Stimulated recall

A stimulated recall is a retrospective interview where a recording of the experiment – e.g. a video – is used as a prompt, assuming that this helps the participants to remember what they were thinking at the time of the event. As such, stimulated recalls are used in combination with other mostly concurrent methodologies, „as a means of triangulation for further exploration“ (Gass & Mackey, 2017, p. 16).

The assumption behind stimulated recalls is that by providing participants with a prompt that requires them to reflect on an event, the access to mental processes during the event is enhanced. The time interval between the event and the stimulated recall should be short to guarantee reliable recalls since „immediately after the task is completed, there remain retrieval cues in short-term memory that allow effective retrieval of the sequence of thoughts“ (Ericsson & Simon, 1996, p. xvi).

The use of videos of eye-tracking experiments as a stimulus for retrospective recalls is controversial. While Dörnyei (2007) recommends providing participants with „rich contextual information”, Gass and Mackey (2017) on the other hand argue that some prompts might be to „noisy” for participants and therefore do not make a good stimulus, like e.g. the video recording of an eye-tracking experiment where participants have to follow their own eye movements on the screen.

Another aspect to consider is the language used for the stimulated recall. If participants are not using their L1, their L2 proficiency might be a problem for performing the task, i.e. understanding the questions they are being asked and expressing themselves. Additionally, the researcher then is challenged with interpreting what participants said (Gass & Mackey, 2017, p. 48).

Information about how these issues were addressed in the current study will be described in the context of the data collection procedure (see section 2.3.6).

2.3.4 Participants

19 international study applicants from preparatory language courses at a large University in Germany took part in this process-oriented study. The sample was part of a larger group of examinees that participated in a field test for the digital TestDaF (see section 1.5.1 for an overview of the research design).

Table 2.1 gives an overview of the participant information. The sample was very similar to the expected test-taker population of the digital TestDaF based on their regional distribution, age and study experience. There were nine female and ten male participants from different nationalities, large groups coming from countries of the Middle East (36,8%), Africa (21,1%) and Asia (21%). For the vast majority (78,9%) German was the L3, English being the first foreign language for many of them. Their ages ranged between 17 and 33 (*M*=24.47; *Standard Deviation (SD)*=3.95). Different fields of study were represented, including business and economics, engineering and humanities, and many participants had already earned a degree in their country of origin: Eight of them had a B.A., four even an M.A., and one was a teacher graduate.

Table 2.1 *Participant information*

Participant	Sex	Age	Country of Origin	Subject	Degree
1-02	F	24	USA	Psychology	B.A.
1-03	M	26	Togo	International Relations	M.A.
1-04	M	27	Turkey	Physics	B.A.
1-05	F	26	Iran	Architecture	B.A.
1-06	M	25	Syria	n.a.	none
1-07	M	20	China	Engineering	none
1-09	M	23	Vietnam	International Relations	B.A.
1-10	F	20	Cameroon	Computer Science	none
2-01	F	23	China	Finances	B.A.
2-02	F	18	Cameroon	n.a.	none
2-03	F	29	Iran	Computer Science	M.A.
2-04	F	32	Iran	Psychology	M.A.
2-05	M	26	Vietnam	Finances	B.A.
2-06	M	25	Cameroon	Physics	Teacher Training
2-07	F	22	Syria	Environmental Engineering	none
2-08	M	23	Uzbekistan	Management	B.A.
2-09	M	20	Russia	German Studies	none
2-10	F	33	Iran	Finances	M.A.
2-11	M	23	Iran	Electrical Engineering	B.A.

Eighteen of the participants took language classes to prepare for the paper-based version of the TestDaF to gain language admission to HE in Germany. None of them had ever taken a computer-based test before. Data on computer familiarity were not collected, but participants had to do a typing test that offered insights on their typing skills (see Section 1.3 for a general discussion of effects on the writing mode, and the following section on the specific typing test used in this study)

2.3.5 Instruments

Integrated writing task

Participants worked on two different versions of the digital TestDaF integrated writing task as described in section 1.2. They were randomly assigned to the two versions (Set 1 and Set 2) of this task: eight participants worked on Set 1, while 11 participants completed the task of Set 2.[21]

The tasks were developed on the basis of existing test specifications to ensure a high degree of comparability. Even though source text features like text length, number of sentences or average sentence or word length seemed comparable, there were still some differences evident between the two tasks, as Table 2.2. shows.

Table 2.2 *Comparison of two test versions of the integrated writing task*

	Set 1	Set 2
Topic domain	Social sciences	Natural sciences
Length of source text	259 words	230 words[22]
Number of sentences in source text	13	13
Mean length per sentence	19.9 words	17.6 words
Mean length of words	5.9 characters	6.4 characters
Flesch reading ease score	53	42
Graph type	pie chart, five data points	bar chart, four data points

The tasks covered different topic domains and used different graph types, also the Flesch reading ease score varied across set.

[21] The tasks are not publicly available due to test security.
[22] The number of words for the source text was increased and determined to 250-300 words (see section 1.2) after early piloting.

This readability index usually ranges between 0 and 100, with higher scores indicating that texts are easier to read.[23] A score of 53 for Set1 is interpreted as fairly difficult to read, while a score of 42 for Set 2 is interpreted as difficult to read.

C-test

Participants also completed a C-test (see 1.5.2) to measure their overall language competence.

Typing test

An additional 2-minute typing speed test for German[24] was used to investigate the influence of the writing medium on the writing process. In this test, participants had to re-type a text that appeared on their computer screen in a text box beneath. The text was the same for every participant and included special features of German like the Umlaut (*ä*, *ö*, *ü*) or the special character *ß*.

2.3.6 Data collection

The data collection included two stages (see Table 2.3). At a first stage, all 19 participants attended a joint session in a computer lab at the languages center where they were enrolled in preparatory language classes. They signed a consent form beforehand (see Appendix A). During this session, participants completed the C-test, the reading comprehension test and worked on the independent task

[23] For German, this readability index is calculated with the following formula: 180 - ASL - (58,5 * ASW). The average sentence length (ASL) is the number of words in the text divided by the number of sentences, while the average number of syllables per word (ASW) is the number of syllables divided by the number of words.

[24] The typing test can be accessed under the following link: https://www.typingtest.com/. Unfortunately, the German version is not available anymore on the website.

of the writing component of the digital TestDaF under exam conditions. The C-test was used in this strand to relate the writing process of participants to their overall language proficiency. The reading comprehension test and the independent writing task were used in the data analysis of Strand 3 for linking this smaller sample of participants to the rest of the field test population in terms of their integrated writing ability and influencing factors.

Table 2.3 *Summary of data collection*

Data collection stage	Instruments
Stage 1: Joint session of approx. 2 hours (normal test condition)	C-Test Reading comprehension test Independent writing task
Stage 2: Individual sessions of approx. 1,5 hours (eye-tracking experiments)	Background questionnaire Integrated writing task Eye-Movement videos Video recordings (external camera) Stimulated recall interviews videos Typing speed test

Data collection at stage 2 included the participants' performance on the integrated writing task of the digital TestDaF, their typing speed, as well as eye-tracking and stimulated recall data. All data were collected in individual sessions. Participants received general information on the eye-tracking experiment beforehand by e-mail. For example, they were asked not to wear mascara since this „is considered a serious problem for data quality" (Holmqvist et al., 2011, p. 119). Participants in need of visual aids were asked to bring both, glasses and contact lenses, to have the opportunity to change from one to the other, in case e.g. the infrared reflection in the glasses turned out to be problematic for tracking eye movements correctly. On site, participants firstly answered questions to collect some personnel data like age, L1, or their fields of study. They then

received background information on the writing task and the purpose of the study. In addition, the experimental set-up in the room was explained. Participants were given the possibility to ask questions if anything was unclear.

Afterwards they completed the summarization task of the digital TestDaF while their eye movements were recorded. After completing the task, the gaze replay of participants' eye movements was used for immediate stimulated recalls. Participants finally completed a typing speed test. Each individual session lasted around 1.5 hours.

Participants were seated across from the researcher at an extra table, approximately 60 cm away from a screen (screen size: 22"; resolution: 1920x1080 pixels) that was connected to the researcher's laptop. A remote SMI eye-tracker was mounted to the center of the lower frame of the participants' monitor to reduce the risk that participants get distracted. Eye movements from both eyes (*binocular eye-tracking*) were recorded as a „screen recording" element in the SMI Experiment center (version 3.7.60) at a sampling rate of 120 Hz. Participants used a mouse and typed on a keyboard with a German layout (QWERTZ).[25] Before starting to work on the integrated writing task, eye movements were calibrated with a built-in 5-point-calibration.

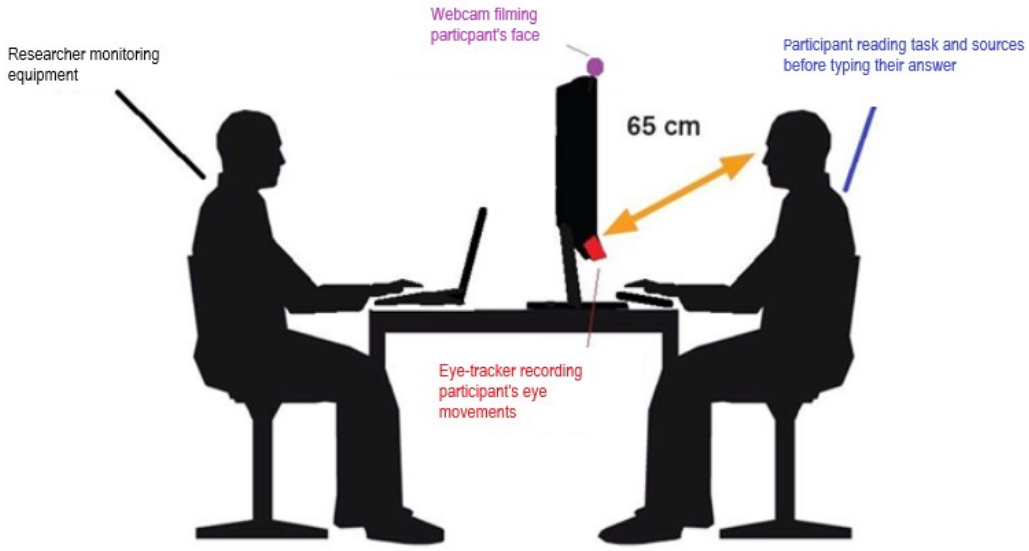
The set-up with an external monitor allowed the researcher to oversee the whole experiment and to observe the eye movements and the writing process in order to make notes for the stimulated recall, e.g. to note down a certain time when a participant was pausing for a long time. Video recordings of an external camera provided more evidence about participants' behavior, e.g. if they looked down at the

[25] Holmqvist et al. (2011, p. 40) caution that typing and/or mouse movements by participants can affect the quality of eye-tracking data. The effect is stronger for high-sensitive recordings (e.g. microsaccades) which were not used in the current study. Therefore, and due to practical constraints, the eye tracker was placed on the same table together with the keyboard and mouse.

keyboard when no eye movements were recorded, to inform the stimulated recall sessions and the interpretation of eye movement data.

Participants had 30 minutes to complete the task. Most of them took up the whole time, only one participant finished early after around 25 minutes.

Figure 2-3 *Set-up eye-tracking experiment*[26]



Immediately after completing the task, each participant took part in a retrospective interview. Even though the language competence of participants is one crucial element in verbal protocols (see section 2.3.3), stimulated recalls in this study were conducted in German due to the very heterogeneous L1-backgrounds of participants.

The gaze replay of the eye movements was used as a stimulus for investigating the cognitive processes involved in their writing

[26] Courtesy of Dr. Nicola Latimer, CRELLA, University of Bedfordshire, UK. Permission to use granted.

process. In order to address the issue of the suitability of the prompt (see section 2.3.3), the "noise" of the gaze replay, i.e. the eye-tracking recording, was reduced by showing fewer data points to the participants. Another option would have been to train the participants in the method of stimulated recall with a corresponding prompt, but this was not possible due to practicality constraints.

Each stimulated recall session followed a certain structure (see Appendix B). The researcher first asked for general impressions, participants were then asked about their text and graph comprehension, i.e. if they were able to summarize the relevant information with respect to the given question orally. Then the researcher selected parts of the video as a stimulus for discussion since it was not possible to watch the gaze replays in full length due to time constraints. Participants were also encouraged to go back to certain episodes of the writing process they remembered as important.

2.3.7 Data analysis

Due to technical reasons during the data collection phase, the data set is not complete for all 19 participants. For example, not all stimulated recall interviews were recorded in full length due to technical problems with the microphone at the beginning of the data collection phase. The review of the eye-tracking data for accuracy also revealed that data from two participants had to be excluded from the data analysis: One participant struggled with the digital test environment over a longer period at the beginning of the recording. He had unintentionally highlighted the whole source text instead of only selecting single words or phrases, and tried to undo this action. His viewing behavior, and hence his writing process were substantially affected by this incidence. For another participant, eye

movements were not recorded accurately during the whole time, even though the calibration beforehand yielded satisfactory results. Different aspects might explain this imprecise measurement: For one thing, the participant frequently changed her seating position during the 30 minutes of task completion. On the other hand, she constantly moved her head to look down on the keyboard. This resulted in the loss of the recordings of her eye movements over a longer period of time. The eye-tracking data of he remaining 17 participants could be included in the data analysis. The researcher decided to remove only the two above mentioned participants from the whole data analysis to have as many data points as possible for each of the data collection stages. Table 2.4 displays the amount of data sets that were used for data analysis in Strand 1.

Table 2.4 *Overview data set Strand 1*

Data collection	Data size used for analysis
eye-tracking	17 recordings
retrospective interviews	14 recordings & transcripts
typing speed test	16 test results

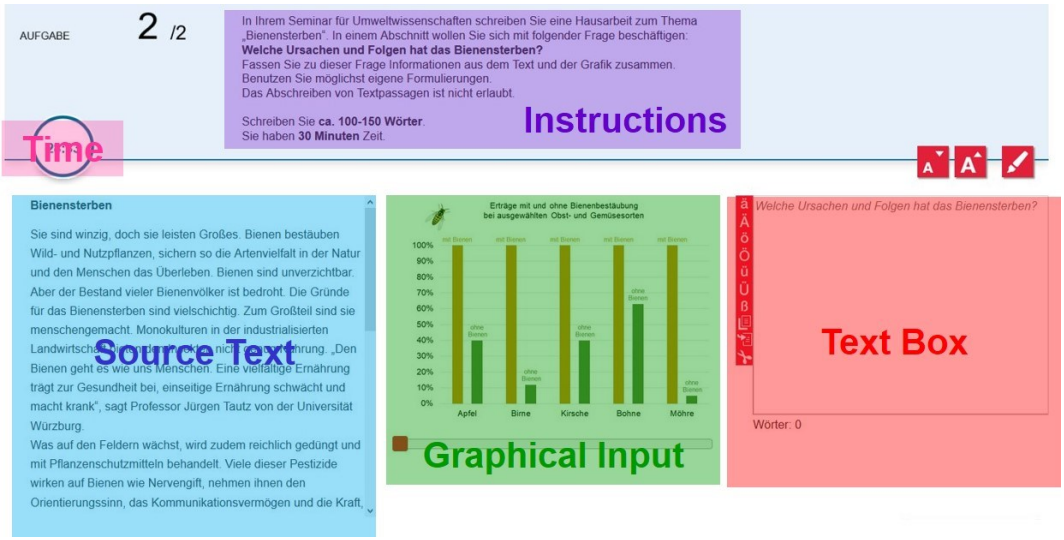
Eye-tracking data

The analysis of the quantitative eye-tracking data should reveal the viewing behavior of participants in relation to different parts on the screen. For this purpose, five so-called *AOIs* were defined in advance in *SMI BeGaze* (version 3.7.42), an analysis software for eye-tracking data.

The definition of the AOIs was based on the assumption that they were relevant for task completion and therefore should play a crucial role in the writing process. As displayed in Figure 2-4, these five

AOIs were: (1) the instructions, (2) the source text, (2) the graphical input, (4) the textbox where participants had to type their text, and (5) the timer.

Figure 2-4 *AOIs for the TestDaF integrated writing task*



These five AOIs covered 40 % of the screen, and the size of the individual AOIs varied across participants (see The different sizes of the AOIs, especially noticeable for the instructions, can be explained by the fact that the instructions in Set 1 and Set 2 were different according to their length, and some participants made use of the possibility to enlarge the font size in the digital test environment. This led to a different display of the AOIs on the screen. Although defined in advance, the AOIs had to be manually edited for every single participant after data collection.

Table 2.5). The input material participants had to relate to during their writing process, i.e. the source text and the graphical input, on average had the largest proportion of all AOIs of the overall screen size (11.28 % for the source text and 10.28 % for the graphical). Smaller amounts were covered by the text box (9.26 %)

and the instructions (8.32 %). In comparison to the other AOIs, the timer only covered a very small area on the screen (0.30%).[27]

The different sizes of the AOIs, especially noticeable for the instructions, can be explained by the fact that the instructions in Set 1 and Set 2 were different according to their length, and some participants made use of the possibility to enlarge the font size in the digital test environment. This led to a different display of the AOIs on the screen. Although defined in advance, the AOIs had to be manually edited for every single participant after data collection.

Table 2.5 *Proportion of individual AOIs on total screen size across sets*

	Instructions	Source Text	Graphical Input	Text Box	Timer
M	8.32	11.28	10.28	9.26	.30
SD	1.25	.23	.15	.05	.00
Min.	7.30	10.50	10.10	9.20	.30
Max.	11.70	11.50	10.50	9.30	.30

Note. All data are percentages

Because the eyes are believed to move synchronously and slight differences might not be that relevant, monocular eye-tracking is common practice in language (testing) research (Conklin et al., 2018, p. 18). Nevertheless, since the eye-tracking system used in this study allowed for binocular tracking, data was checked again for each of the 17 participants. The validation results in particular showed some major differences between the left and the right eye in terms of deviation from the calibration points (see Appendix C).

[27] Even though recommended (see Holmqvist et al., 2011, p. 206ff.), it was technically not possible to define the remaining area as *whitespace* in the screen recording element of the eye-tracking software.

Hence, only data for each participant's 'best eye' was included in the analysis of all eye-tracking metrics.

The eye-tracking software offers different measures for each individual AOI, but since the length or the size of an AOI determines the measures that should be examined, only few of them were of interest for the research questions.

The current study focuses on the following AOI related eye-tracking measures:

- ***Dwell time***, i.e. how much time was spent in an AOI over the whole trial.

Dwell time is a measure that not only takes into account the duration of all fixations within an AOI (like e.g. *fixation duration*), it is rather the sum of durations for all gaze data sample, i.e. fixations and saccades. A *dwell* on an AOI is something that is usually longer than a single fixation, and it could contain more than one fixation. Dwell time is calculated for each visit to an AOI, starting with the first fixation, and ending with the last fixation within the same AOI. For this study, the proportion of dwells participants spent in the different AOIs during task completion was analyzed by looking (a) at the actual time participants spent in the different AOIs in seconds, and (b) relating the dwell time to the individual processing time, i.e. the time each participant effectively worked on the task.

- ***Transitions***, i.e. the movements from one AOI to another.

For calculating transitions, the eye-tracking software only takes into account fixations. The transitions to/from the individual AOIs were related to the total number of transitions for every participant to take into account expected deviations in the total number of transitions across participants. The average percentage of transitions from one specific AOI to another during task completion across participants is reported in a 2D transition matrix, not taking

into consideration if there have been additional fixations between the fixations on AOIs, for instance, if candidates looked somewhere else on the screen (e.g. on the screen not defined as an AOI like the *whitespace*), or if they took their eyes off the screen to look down on the keyboard.

- ***Revisits***, i.e. the number of returns to an AOI.

Revisits are calculated as the number of dwells returning to a specific AOI from outside, requiring that the participant has looked at least once at the AOI earlier. The number of revisits therefore equals the number of dwells in the AOI minus one (see Holmqvist et al. 2011, p. 423.).

Following Plakans's model of reading-to-write tasks (Plakans, 2008), *dwell time* and *transitions* were calculated for two different stages during the writing process – pre-writing and writing – in order to answer RQ1c. Pre-writing was defined as the interval between the beginning of task completion and the first keyboard entry of each participant, while the writing stage encompasses the time from the first keyboard entry to the end of task completion. Pre-writing and writing time were determined by looking at the gaze replays of the eye-tracking recordings.

All data was analyzed for the whole sample (N=17), as well as for the two different sets participants were randomly assigned to (Set 1 and Set 2) to look into possible task effects.

Stimulated recalls

The 14 retrospective interviews that were included in the analysis lasted between 16:21 minutes and 28:50 minutes (*M*=25:08 min; *SD*=3:14 min). The recordings were transcribed verbatim. To get faster access to the content, the transcription of the retrospective

interviews basically followed the conventions for simple transcripts (Dresing & Pehl, 2018). But unlike simple transcripts, the transcripts in this study also included some non-verbal and procedural information.

The following transcription conventions were applied (adapted from Kuckartz, 2016; Barkaoui, 2016):

- The utterances of the participants were not transformed into written German and were not corrected, i.e. grammatical structures and wording were maintained in the transcription, even if they were wrong.
- Pauses, even significant long pauses, were not marked.
- Affirmative utterances (like *mhm, ja, aha*) were transcribed, as well as fillers (like *äh*, *uhm*).
- Words with a special emphasis are CAPITALIZED.
- Non-verbal activities and procedural behaviors are noted in double brackets, e.g. ((points to the screen)), ((laughs)).
- Incomprehensible words are indicated by (inc).
- Text that interviewees read from the tasks is not included for test security reasons. These passages are marked by [XXX].

The transcripts were coded in NVivo 12 Plus. In a first coding cycle, deductive and inductive coding techniques were adopted using the Simultaneous Coding method (Saldaña, 2016). For relating the qualitative interview data with the quantitative eye-tracking data, and in order to answer the research questions stated in section 2.3.1., segments of the transcripts that were determined to be conceptually meaningful were coded according to the following multiple dimensions:

- stage of the writing process, i.e. pre-writing or writing
- the AOI participants were referring to

- type of eye movement, e.g. whether participants commented on long dwells in an AOI for a longer time, or whether they were referring to parts in the writing process where their eyes moved between the different AOIs
- the cognitive processes described in this segment
- factors that might have affected the cognitive processing.

At this stage, provisional codes – some already predetermined like the stage of the writing process, the different AOIs, or the eye movement – were complemented using an inductive initial coding approach. At a transitional stage, codes were condensed and redefined by the process of „shop-talking" (Saldaña, 2016, p. 231), i.e. a regular exchange on the data analysis with other peers in a research colloquium. Additionally, code mapping (Saldaña, 2016) served as a basis for categorizing the cognitive processes in processes of reading, writing, shared and test-taking processes based on Plakans, Liao et al. (2019). The interviews were then recoded with the revised coding scheme (see Appendix D).

In order to see whether or not the qualitative data analysis „has meaning that extends beyond an individual researcher" (O'Connor & Joffe, 2020, p. 3), a second coder independently coded a subset of three randomly selected interviews after receiving a short familiarization with the coding scheme. Prior coded data served as examples to clarify any ambiguities. After the second coder finished coding, a coding comparison was run in NVivo to check inter-coder reliability (ICR) in terms of percentage agreement and a Kappa coefficient. In NVivo, the percentage agreement for sources that are character-based (like the transcripts of the stimulated recalls) is the sum of content coded to a selected node by both coder A and coder B and the content not coded by either coder. For example: Two coders coded a source with 1000 characters. They assigned the same 200 characters to a code, 100 characters were only coded by coder

B, and the remaining 700 characters were not coded by either coder. The percentage agreement then would be (200 + 700) / 1000 = 90%.

Results for the average ICR on the primary code level are displayed in Table 2.6. The entire agreement table can be seen in Appendix E).

As can be seen, the percentage agreement for the two coders is high, ranging from 87.62% for the coding of eye movements and 96.19% for coding of the cognitive processes. With 80-90% as a minimal benchmark for percentage of agreement between the two coders (Saldaña, 2016), the results show that the two coders highly agreed on their coding decisions.

Table 2.6 *ICR for primary codes across double coded interviews*

Primary Code	Agreement (%)	Cohen's Kappa
Phase	93.10	.85
AOI	90.91	.59
Eye movement	87.62	.43
Cognitive processes	96.19	.47
Influencing variables	92.40	.37

Cohen's Kappa coefficients show fair to excellent agreement for most of the codes, while ICR for coding the influencing variables was poor (.37).[28] Coding reliability was highest for *phase* and the lowest for influencing variables. The reason why a high percentage agreement is not necessarily associated with a high Cohen's Kappa coefficient, is mainly related to two different aspects:

[28] According to the NVivo manual, Cohen's Kappa should be interpreted as follows: values below .40 show a poor agreement; .40 – .75, fair to good agreement; over 0.75, excellent agreement.

For one thing, studies have shown that ICR tends to be lower when more codes are available (Hruschka et al., 2004). There were only two categories to distinguish for *phase* compared to five or more categories for the other codes, which might have affected coding reliability. However, the relatively low Kappa coefficients are mainly caused by the fact that automated calculated ICR like in NVivo depend on the selected text units for coding. In this study, no pre-specified small data units like sentences or passages were used as a basis for coding, but coders assigned codes to longer ad hoc segments that were conceptually meaningful to them. While longer data units typically increase the validity of interpreting the data, they often lead to poorer reliability. If the two coders differ in the selection of text segments they assign to a certain code simply by a digit or a punctuation mark, ICR can be compromised since software packages like NVivo calculate Cohen's Kappa on the character level (O'Connor & Joffe, 2020; Kim et al., 2016). A suggested correction for the biased Cohen's Kappa in NVivo by Kim et al. (2016) was not adopted for this study because it involved a binary coding approach. Instead, codings were reviewed again to check whether the two coders applied the same codes instead of focusing on start/end of a coded segment, as suggested by O'Connor and Joffe (2020).

Since this was the case and given the limitations of Cohen's Kappa in NVivo, Kappa values in this study were interpreted in conjunction with the percentage agreement as satisfactory.

Typing test

Besides the C-test which was used to measure participants' overall language competence, typing test results reported different parameters, but mainly assigned the participants to one of five

categories, i.e. slow, average, fluent, fast and pro typist, according to their adjusted typing speed (Figure 2-5)

This adjusted speed is based on the typing speed and the accuracy of the typing. Typing speed takes into account how many words per minute (WPM) a participant typed and the total number of chars he/she produced during the two minutes of the test. For the accuracy of typing, the number of mistyped words was counted.

Figure 2-5 *Example report for typing test results*

2.4 Findings[29]

This section will report findings in relation to the overall research aim of this process-oriented strand, namely: What are the cognitive processes that test takers use when working on the summary writing task of the digital TestDaF?

Since most findings draw on the quantitative and the qualitative data, results will not be reported separately. Instead, findings from the eye-tracking experiment, the C-test and the typing test will be

[29] Some results were previously published in Zimmermann (2020a; 2020b).

presented in combination with insights from the stimulated recalls to answer the different sub-questions.

2.4.1 Approach to the task

The first research question (RQ 1a) was to find out how test takers approached the task. To inform the answer to this research question, the individual time for pre-writing and writing was determined for each participant.

Pre-writing and writing time

As Table 2.7 shows, the average pre-writing time, i.e. the time before participants typed their first letter on the keyboard, was a little over 4 minutes long, while the time they took for writing was on average around 25 minutes.

Table 2.7 *Average pre-writing and writing time*

	pre-writing	**writing**
M (SD)	04:11 (01:36)	25:32 (01:44)
Mdn	04:08	25:39
Min.	00:25	22:32
Max.	06:27	29:35

The way participants approached the task varied with regard to the time they spent before starting writing and actually writing during task completion. While some participants spent over six minutes on their pre-writing time, others started writing almost right away.

Some participants commented on their individual approach, like participant 2-01, who had the shortest pre-writing time with only 25 seconds. She decided not to process the text as a whole before she started to write, instead she chose to read short passages of the text and then to summarize the information she just read:

> Ich möchte Zeit sparen. Äh, wichtiger Grund ist, ich lesen die ganz Text und ich werde vergessen und es ist nicht nützlich für mich, die ganze Text zu lesen und zu verstehen, und dann schreiben. [*I wanted to save time. One important reason is that I will forget if I read the whole text first, and it is not useful for me to read and understand the whole text and then start to write.*]

This approach of summarizing sentence by sentence is highly individual; as all other participants processed the sources first before starting to write. The average pre-writing time of about four minutes was on the one hand related to reading since participants needed a certain amount of time for processing the input material. Some participants reported on difficulties comprehending the source text (see also section 2.4.4 on the effect of overall language competence), therefore spending more time looking at the source text, re-reading it in order to understand unfamiliar words, hence increasing the pre-writing time automatically. On the other hand, some participants were also engaged in extensive planning processes during pre-writing.

2.4.2 Engagement with the task and reading-writing-relations

The focus of the second research question (RQ 1b) was to find out about test takers' engagement with task, and to examine whether the processes involved are more related to reading or writing.

To answer this research question, the dwell time in the AOIs, as well as the transitions between and the revisits to the different AOIs

were analyzed. Codings from the stimulated recalls were used to examine the underlying cognitive processes participants engaged in during task completion.

Dwell time

It is apparent from Table 2.8 that the text box had the highest dwell time of all AOIs with around nine minutes, followed by the source text with approximately seven minutes. Adding up all the AOIs that are related to input material (i.e. the source text, the graphical input as well as the instructions), participants on average spent around 33% of the total task duration in these AOIs, a little more than they spent looking at the text box (nearly 30%). In comparison, the amount of dwell time for the remaining time to work on the task was quite small. Overall, participants spent around 63% of the total task duration looking in the AOIs.

The external video recordings showed that for most of the remaining time participants' eyes were not directed towards the screen, they were rather looking down at the keyboard. The high dwell time in the source text can be explained by the fact that participants had to read and comprehend the source text in order to work on the task. Some participants focused on unfamiliar words for a longer time in order to figure out the meaning. Participant 2-03 for example said the following when asked why he looked at a certain word in the source text:

> Ich verstehe nicht diese Adjektiv. Ja, und was ist die ähnliche Wörter für diese Adjektive. Und ich recherchiere in meinem Gehirn. [*I don't understand this adjective. Yes, and what are similar words to this adjective. And I searched my brain.*]

To understand the source text, participants occasionally even had to re-read parts or the whole text more than once. A few

participants voiced that this was related to a lack of concentration, like participant 1-02:

> Ich konnte mich nicht konzentrieren zum Anfang den Text zu lesen. Ich erinnere mich, ich habe vielmal, ich würde anfangen, und dann zurückgehen, und dann anfangen. Ich habe gemerkt, dass ich habe nicht verstanden, was ich gelesen habe. [*I couldn't concentrate at the beginning to read the text. I remember that I often started all over again. I noticed that I hadn't understood what I had read.*]

Participant 1-04 also had problems concentrating and blamed the setting for the need to re-read the text again at a later stage of her writing process:

> Ja, äh, das war die Probleme, ich habe Ihnen gesagt. Ich erinnere mich mit, äh, wie kann man das sagen, ich erinnere mich nichts wegen Stress, ja ich weiß, dass das nicht Original-Prüfung. Aber ich habe Stress irgendwo hier. [*Yes, that was the problem I told you. I can't remember how to say that, I can't remember anything because I was stressed. I know that this is not a real test, but I'm stressed anyway.*]

While the high dwell time in the source text was mainly related to reading, the dwell time in the text box on the other hand was primarily related to writing processes, including reading for revision. When looking at the text box, participants either looked at this AOI to see their writing on the screen, or they were re-reading their own text critically for evaluation and revision. At certain points in the revision process, participants paused in order to think about rephrasing what they had written, as participant 1-03 pointed out:

> Ja, möchte ich, wollte ich eine Vergleichung machen zwischen diesen Teil und den anderen. Aber ich habe gedacht, ne, ein Vergleichung passt nicht. Ich wollte, es, ein Wort fehlt mir, das Wort ist "gefolgt von". Ja, ich habe lang gedauert, das, das war ein bisschen, ich habe, ich habe das vergesst, und das nimmt viel Zeit. [*Yes, I wanted to compare this part here and that one. But I thought, no, a comparison does not fit here. I wanted, a word was missing, the word "followed by". Yes, it took long, that was a little, I had forgotten, and that took a long time.*]

Despite the fact that the time to work on the task was limited, participants regarded revision as key and incorporated that process

into their writing. Especially towards the end of the 30 minutes task duration, participants mostly looked in the text box, reading their own text again in order to correct mistakes, as participant 1-09 explained:

> Ja, weil ich, ich denke, es ist die Zeit für Korrigieren. Ich muss noch einmal lesen und dann korrigieren die Fehler. [*Yes, because I think it's time for revising. I need to read again, and then correct the mistakes.*]

As Table 2.8 also reveals, viewing behavior of participants varied to a great extent: While some participants spent almost no time looking at the instructions, others took more than two minutes to read them during task completion. The dwell time for the graphical input also varied significantly between participants, ranging from around eight seconds to almost four minutes during the 30 minutes of task completion. The time participants paid attention to the remaining time also differed. All participants spent at least four minutes in the AOIs of source text and text box.

Beyond the time participants spent overall in the different AOIs, it is worth looking at how often they went back to the individual AOIs during task completion.

Table 2.8 *Average dwell time in different AOIs during task completion*

	Instructions		**Source Text**		**Graphical Input**		**Text Box**		**Time**	
	in seconds	% of total task duration	in seconds	% of total task duration	in seconds	% of total task duration	in seconds	% of total task duration	in seconds	% of total task duration
M (SD)	56.13 (40.10)	3.03 (2.18)	422.85 (125.51)	23.17 (6.37)	124.20 (62.07)	6.79 (3.38)	547.60 (208.16)	29.97 (11.16)	6.79 (4.48)	.37 (.26)
Mdn	60.80	3.30	419.85	23.00	116.65	6.20	531.33	29.20	5.64	.30
Min.	.53	.00	242.16	13.20	8.21	.40	248.98	13.60	1.96	.10
Max.	143.99	7.80	636.74	34.30	236.56	13.00	847.70	46.20	14.38	.80

Note. N=17.

Revisits

Participants revisited the different AOIs in varying degrees. The overall number of revisits to all AOIs varied across participants, ranging from 156 to 549 (*M*=328.71, *SD*=106.33). Due to these substantial differences, the number of revisits to the individual AOIs are reported in relation to the total number of revisits for every participant.

As Table 2.9 shows, the AOI participants on average revisited most of all was the text box. Revisits to this particular AOI account for 45% of all revisits, followed by the source text (24%) and the graphical input (around 21%). Revisits to the input material make up as much as the revisits to the text box. Less frequently participants went back to the instructions or checked the remaining time. The percentage of the revisits to each of these AOIs was around 5%. Again, there were individual differences: For example, revisits to the instructions account for 5% of all revisits on average. While some participants never looked at the instructions again during task completion, others revisited this AOI quite often. Similarly, the percentages of revisits to the graphical AOI range from 1% to over 37%.

Revisits are especially informative in relation to the dwell time (see Table 2.8): Whereas the average dwell time in the graphical input was relatively low, participants revisited this AOI quite often. This becomes especially apparent in comparison to the time participants looked at the source text, and the extent to which they revisited this AOI. While the dwell time for the source text was much higher than for the graphical input, the revisits to both AOIs were similar. In the stimulated recalls participants did not touch upon this specific aspect, but the discrepancy between dwell time and revisits for the source text and the graphical input could be an indication that reading texts and visuals are processed differently.

Table 2.9 *Revisits to different AOIs during task completion*

	Instructions		**Source Text**		**Graphical Input**		**Text Box**		**Time**	
	number	% of total revisits	number	% of total revisits	number	% of total revisits	number	% of total revisits	number	% of total revisits
M (SD)	17.12 (9.27)	5.03 (4.58)	97.47 (58.56)	24.00 (11.33)	78.71 (32.86)	20.91 (7.83)	169.29 (45.83)	44.72 (6.00)	20.12 (12.89)	5.33 (3.31)
Mdn	19.00	4.75	97.00	21.89	86.00	19.85	180.00	44.68	21.00	4.74
Min.	0	.00	15	8.85	5	1.28	63	32.46	4	1.51
Max.	32	20.51	197	46.94	130	37.61	231	58.70	40	11.54

Note. N=17

How participants engaged with the input material can best be informed by looking at the transitions between the different AOIs.

Transitions

The overall number of transitions varied across participants, ranging from 96 to 431 (*M*=262.18, *SD*=84.76). Due to these substantial differences, the number of transitions to/from the individual AOIs were set in relation to the total number of transitions for each participant. The average percentage of transitions from one specific AOI to another during task completion across participants is reported in a 2D transition matrix (see Table 2.10).

Table 2.10 *Average percentage of transitions during task completion*

	Instructions	**Source Text**	**Graphical Input**	**Text Box**	**Time**
Instructions		.96 (.76)	.94 (.79)	2.29 (1.90)	.57 (1.76)
Source Text	.99 (.67)		4.37 (2.31)	13.74 (10.69)	2.66 (1.82)
Graphical Input	1.41 (1.73)	4.24 (2.40)		21.10 (9.32)	.38 (.55)
Text Box	1.86 (2.00)	14.30 (10.40)	20.76 (9.72)		1.84 (1.76)
Time	.46 (1.24)	2.62 (1.93)	.78 (1.08)	3.70 (2.64)	

Note. N=17. SD is in parentheses.

All data add up to 100%, with each cell representing the percentage of the specific transitions in relation to the total number of transitions. In this matrix, the AOI in each row indicates the starting point of participants' views, while the columns specify the AOI where their

glance was directed to. For example, the transitions from the source text to the graphical input account for 4.37% of all transitions (third column, second row). Transitions from the graphical input to the text box and vice versa were most frequent with around 21% of all transitions in each direction, followed by transitions between the text box and the source text (approximately 14%). Participants' looks did not frequently move between the two sources, these transitions only account for around 4% of all transitions. The transitions between the two sources and the text box account for over 78% of all transitions.

The stimulated recalls revealed that test takers went back to the input material mainly for the following reasons:

Firstly, they moved between the source text and the graphical input to select key information. Sometimes this required the comparison of both sources, like participant 2-06 explained:

> Ja, ich möchte wissen, welche Zusammenhang es zwischen diesen letzten Satz und diesen ersten Satz hier. Und dann wollte ich wissen, was hat dieser Satz mit der Grafik zu tun? Gibt es Informationen, die, die widersprechen in Grafik? [*Yes, I wanted to know the link between the last sentence here, and the first sentence here. And then I wanted to know if there was a link to the graphical input. Is there any information in the graphic that contradicts this?*]

Occasionally the selection of key information was supported by previously highlighted passages or key words in the sources, but this was not always the case, as participant 1-07 had to admit:

> Meine, meine Markierung funktioniert nicht so gut. Ich kann die Stelle nicht finden. Und ich habe mich beeilt. [*My highlighting didn't work well. I can't find the passage. And I'm in a hurry.*]

Besides selecting key information from the sources, participants also went back to the input material to evaluate their own writing, i.e. to check whether the information they had included in their writing was correct. For example, participant 1-02 went back to the source text after she had read the text written so far to search for other key information:

> Ähm, ja, ich glaube, bei diesem Punkt habe ich, habe ich gemerkt, dass ich muss eine Vorteile geben, nicht nur die Nachteile, denn weil es nicht richtig war. [*Ähm, yes, I think at this point I realized that there had to be advantages, and not only disadvantages. Because this was not correct.*]

Additionally, the sources provided the participants with language material. Even though the instructions state that test takers are not allowed to copy longer passages, the use of key words is still allowed. Participants therefore drew on both sources, either to check the spelling of key words they included in their own text, or to use the language material in the sources as a basis to search for paraphrases or synonyms, like participant 1-02:

> Ja, ich denke, ich hab, ich habe nach dieses Wort gesucht. Oder eine andere. Weil ich kenne dieses Thema nicht so gut. Ich hatte nicht selber gute Ideen zum Wortschatz. Also, ich denke, ich hab es nicht noch mal gelesen, aber ich war schnell am Suchen. [*Yes, I think I was looking for this word, or another. Since I'm not very familiar with the topic, I lacked the vocabulary. So, I think I didn't read it again, but I was scanning the text.*]

Reading-writing relations

In order to see whether the involved cognitive processes are more related to reading or writing, or were so-called "shared processes" (Plakans, Liao et al., 2019), the processes identified in the stimulated recalls were grouped into four different categories with different sub-processes within each category, based on Plakans et al. (2019): (1) writing, (2) reading, (3) shared processes, and (4) test-taking.

Overall, there were 213 references coded in the stimulated recalls for cognitive processes. The proportion of references was nearly equally distributed across the four categories, with 26.92% of the cognitive processes being coded as shared processes, 24.88% as writing processes, 23.94% as reading, and 22.07% as test-taking strategies. A category of "random process" was added since some participants mentioned in the retrospective interviews that their viewing behavior

had no specific purpose, or they were not sure about the underlying process, like e.g. participant 1-07 described:

> Also, das ist, äh, ich finde sinnlos, weil, äh, das ist nur, ich weiß nicht, warum ich noch einmal nach die Frage gucke. [*I find it pointless, because it's just I don't know why I looked at the question again.*]

However, this category only accounted for 2.82% of all cognitive processes.

The coding references, i.e. the number of coded segments and the according sub-processes are listed in Table 2.11. As the table shows, all of the participants whose stimulated recalls were included in the analysis (N=14) engaged in processes of writing and reading as well as in blended reading-writing-processes, almost all of them commented on test-taking strategies.

The writing-related processes comprised planning (*pausing for thinking*), formulating (*finalizing the text*) and revision (*deleting parts of the text*, *revising*) – processes that can also be linked to existing writing models (section 2.1). Taking a closer look at the cognitive processes that test takers commented on in relation to the different AOIs, a Matrix Coding Query in NVivo was run to explore the intersection of the nodes Cognitive Processes and AOI (see Figure 2-6). It is not surprising, that writing processes were linked to the AOI of the text box. Reading on the other hand was mainly related to processing the input material, i.e. the source text and the graphical input. Participants commented on reading in relation to comprehending the source text and the graphical input, and they also mentioned reading strategies like highlighting information. But as displayed in Table 2.11, reading-related processes were also subsumed under writing (*reading one's own text for revision*), test-taking strategies (*reading to define the task*), and under shared processes. Reading processes that were subsumed under the processes of reading and writing were mostly related to dwells in an AOI. As already mentioned above, participants'

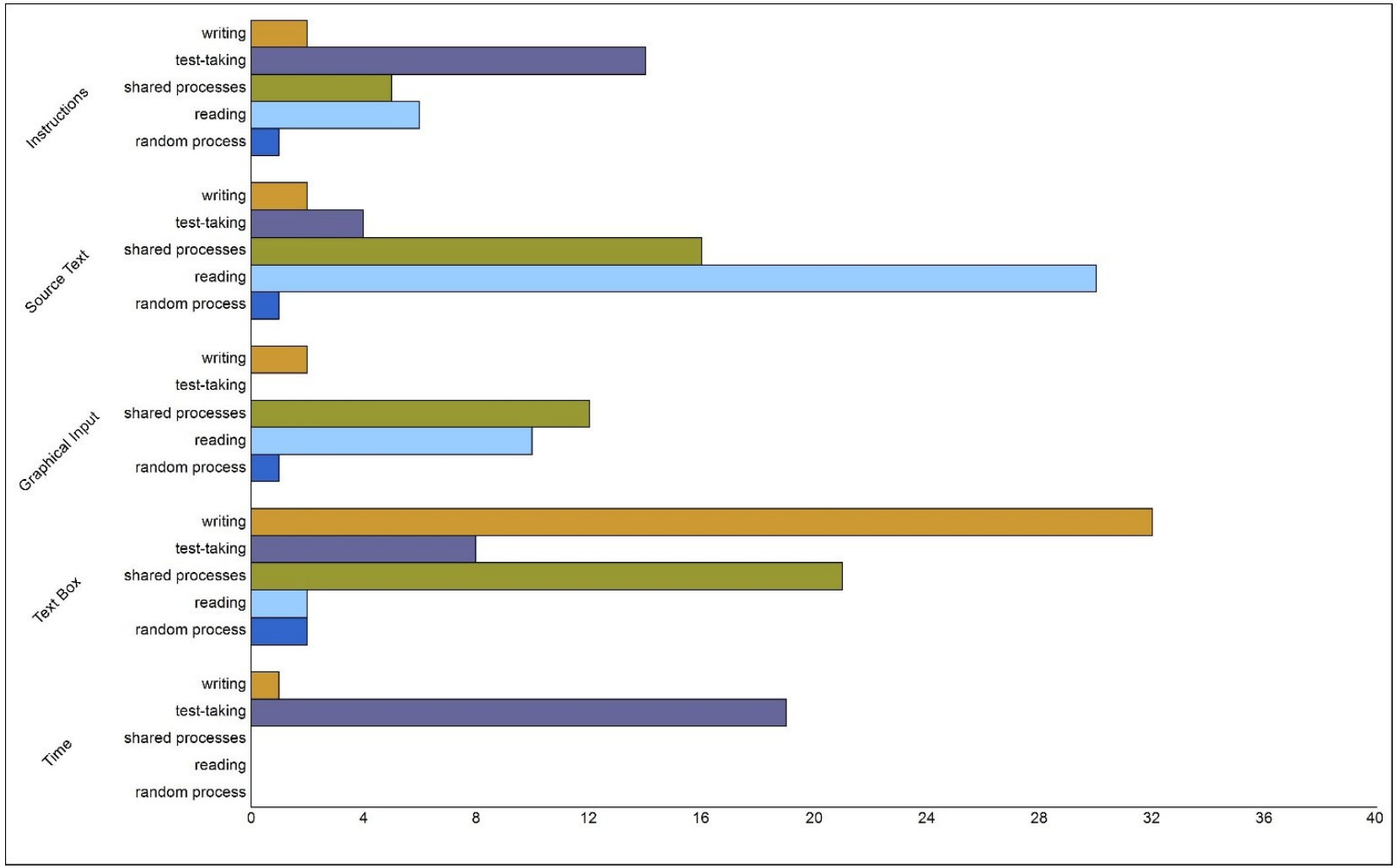
high dwell time in the AOI of source text and text box can be explained by the fact that they were either reading the source text for comprehension, or they were reading their own text for revision. On the other hand, reading processes that were subsumed under shared processes were more related to transitions between the different AOIs.

Table 2.11 *Coding references for cognitive processes*

Cognitive processes	**Participants**	**Total references**
shared processes	**14**	**56**
looking for ways to express something	14	23
identifying connection between text and graphical input	7	11
evaluating own writing	8	9
returning to highlighted passages	6	7
checking spelling of keywords in the sources	4	5
summarizing sentence by sentence	1	3
writing	**14**	**53**
finalizing the text	9	12
deleting parts of the text	7	11
reading one's own text for revision	9	10
revising	7	8
pausing for thinking	7	8
formatting	2	3
inserting a placeholder	1	1
reading	**14**	**51**
reading source text for comprehension	12	19
processing graphical input	8	10
highlighting information	8	10
understanding the meaning of unfamiliar words	7	8
selecting information from the source text	3	4
test-taking	**13**	**47**
time management	12	24
reading instructions to define the task	10	18
checking word count	4	5
random process	**5**	**6**

As Figure 2-6 shows, shared processes were mentioned by the participants most frequently in relation to the AOI of text box and the input material, i.e. the source text and the graphical input. And as the eye-tracking data already have revealed, the transitions between the text box and the source material were the most frequent (Table 2.10). These 'integrated' processes comprised strategies where test takers either went back to the sources to draw on existing language material for their own writing, or to evaluate their own writing by checking the information in the sources to see whether they included them correctly in their text, sometimes returning to prior highlighted passages. This evaluation of one's own writing and the according reading-writing-processes involved were subsumed under shared processes and not coded as writing processes, because they were different from *reading for revision*, where test takers only focused on their own writing in order to correct the grammatical mistakes. Overall, eye-tracking data and insights from the stimulated recalls revealed that participants engaged with the task as predefined by the task format and the instructions, i.e. processing the input material and writing their own summary based on the information presented in the source text and the graphical input. Cognitive processes involved reading and writing processes with according sub-processes, but participants also used shared processes, i.e. processes that comprise specific reading-writing-processes. Since participants worked on the integrated writing task under test conditions, they also employed test-taking strategies, mainly related to reading the instructions and checking the remaining time.

Figure 2-6 *Cognitive processes in relation to the AOIs*



2.4.3 Cognitive processes at different stages of the writing process

The third sub-question was: Do the cognitive processes vary in relation to different stages of the writing process?

In order to answer that question, the time participants spent in the different AOIs, as well as the transitions between the AOIs were related to the individual time participants spent on pre-writing and writing (see section 2.4.1).

Dwell time

By breaking down the time participants spent in AOIs for distinct stages of the writing process, one can see that viewing behavior differed for pre-writing and writing time (see Table 2.12). During pre-writing participants spent most of their time looking at the source text (almost three minutes on average), reading the instruction (30 seconds) and processing the graphical input (nearly 25 seconds). They almost spent no time looking at the remaining time during pre-writing, but did so during writing. This also applies for the AOI text box: from the total time participants dwelled into the text box, almost all of that dwell time occurred during writing.

Even though participants spent most of their pre-writing time looking at the source text, the amount of time they spent in this particular AOI was even higher during writing (almost four minutes on average). The dwell time for processing the graphical input was also higher during writing (almost 100 seconds) than during pre-writing, whereas the dwell time for reading the instructions was equal for both stages of the writing process. Again, differences across participants are noticeable.

Table 2.12 *Average dwell time in different AOIs during pre-writing and writing*

		Instructions		**Source Text**		**Graphical Input**		**Text Box**		**Time**	
		in seconds	%	in seconds	%	in seconds	%	in seconds	%	in seconds	%
Pre-Writing	*M (SD)*	30.19 (26.90)	10.40 (7.59)	162.27 (60.76)	57.88 (15.07)	24.95 (13.86)	8.82 (3.92)	6.68 (6.78)	2.90 (3.32)	.73 (.64)	.29 (.27)
	Mdn	23.03	10.42	180.67	58.93	27.76	8.10	4.81	1.74	.52	.19
	Min.	.00	.00	4.33	9.61	.99	2.21	.19	.08	.00	.00
	Max.	94.28	23.16	255.45	75.25	60.59	13.00	26.94	14.00	2.57	1.08
Writing	*M (SD)*	25.83 (23.00)	1.65 (1.42)	260.58 (128.56)	16.66 (8.03)	99.26 (57.90)	6.40 (3.66)	540.92 (201.10)	34.98 (13.60)	6.06 (4.09)	.39 (.25)
	Mdn	20.28	1.42	285.50	17.16	94.52	5.93	525.94	32.23	4.91	.34
	Min.	.30	.02	7.15	.50	7.22	.40	247.47	17.39	1.62	.12
	Max.	83.57	5.06	442.75	28.94	205.85	12.47	830.36	57.00	13.53	.87

Note. N=17. % is the percentage of dwell time related to the writing stage, i.e. pre-writing and writing time.

Transitions

In a next step, transitions were analyzed in relation to the distinct stages of the writing process. The results are displayed in Table 2.13.

The percentages are now related to the total amount of transitions for each phase, i.e. pre-writing and writing. For example, the transitions from the graphical input to the instructions during pre-writing account for 3.32% of all transitions during pre-writing.

During pre-writing participants most frequently transited between the source text and the graphical input (16%) and reversely (almost 15%). A high proportion of transitions can also be noticed from the graphical input to the text box (13%), which might be surprising since participants had not started writing their text at this stage. But it is worth pointing out in this context again that the text box included the question that participants had to answer, i.e. to synthesize information from both sources related to this question (see section 1.2). Apparently participants also checked the remaining time while reading the source text before starting to write, as the percentage of transitions between these two AOIs, i.e. the source text and the time, is relatively high with approximately 7 % of all transitions during pre-writing.

While writing, participants most frequently looked from the text box to the graphical input and from there back to their own writing (22%), followed by transitions from the text box to the source text and vice versa (around 15%). Transitions between the two sources, i.e. the source text and the graphical input, which have been frequent during pre-writing decreased during writing, accounting for only 3 % of all transitions during the writing stage. The proportion of looks from the text box to the remaining time and back to the text box on the other hand increased clearly during writing; transitions

Table 2.13 *Average percentage of transitions during different stages of the writing process*

	Instructions		**Source Text**		**Graphical Input**		**Text Box**		**Time**	
	Pre-Writing	Writing	Pre-Writing	Writing	Pre-Writing	Writing	Pre-Writing	Writing	Pre-Writing	Writing
Instructions			4.65 (4.98)	.60 (.57)	2.53 (4.32)	.76 (.67)	4.85 (8.86)	2.30 (2.39)	2.49 (4.96)	.44 (1.59)
Source Text	2.74 (4.31)	.74 (.45)			16.53 (10.45)	3.36 (2.17)	2.58 (4.12)	14.59 (11.30)	7.24 (5.43)	2.30 (1.92)
Graphical Input	3.32 (5.97)	1.29 (1.82)	14.57 (16.91)	3.51 (1.97)			13.61 (8.98)	21.58 (9.71)	.00 (.00)	.41 (.58)
Text Box	3.81 (4.94)	1.60 (1.66)	4.17 (5.20)	15.08 (10.78)	5.51 (5.23)	22.00 (9.96)			.23 (.93)	2.00 (1.95)
Time	1.88 (4.65)	.34 (.95)	7.96 (6.68)	2.22 (1.96)	1.17 (2.83)	.84 (1.31)	.15 (.64)	4.02 (2.82)		

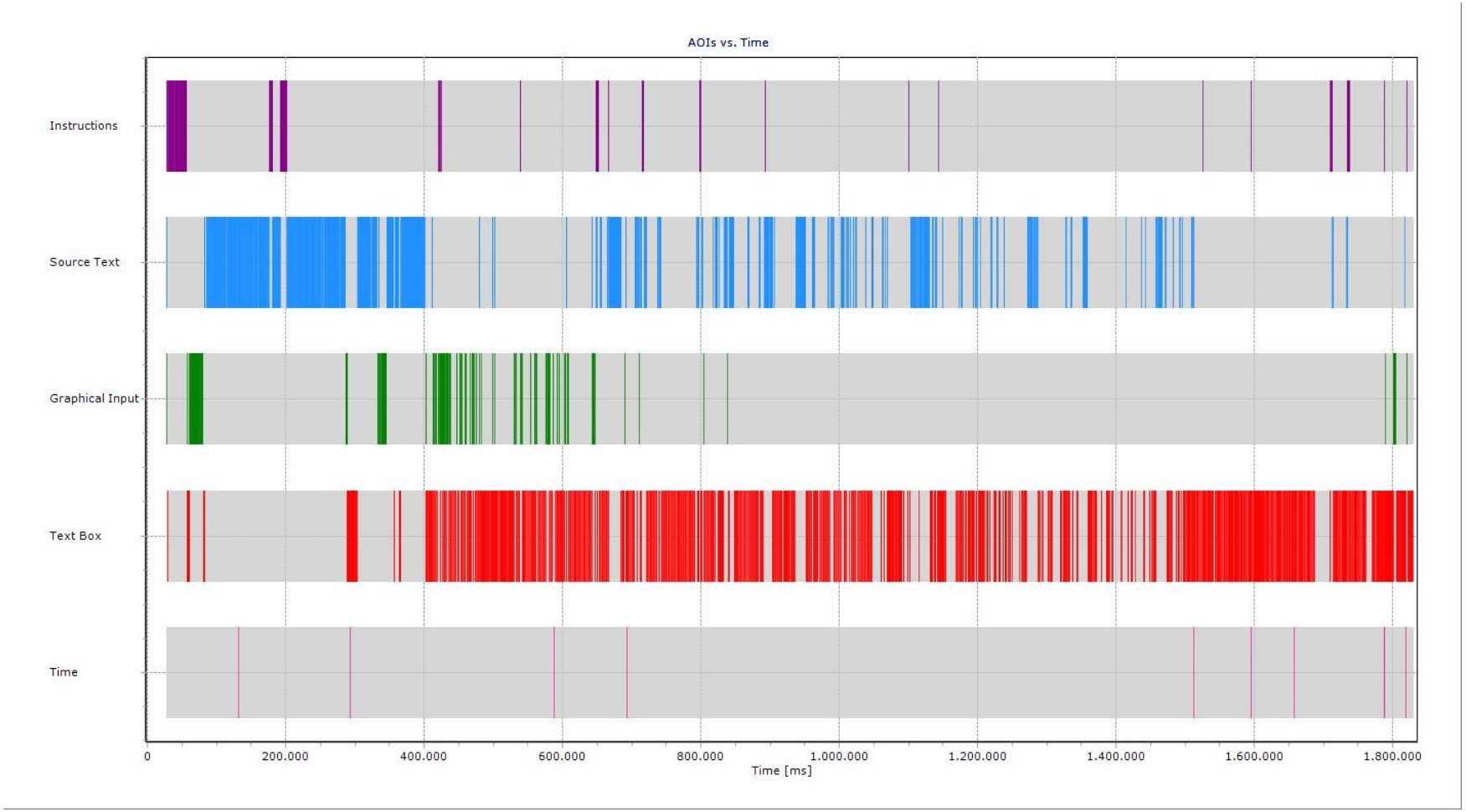
Note. N=17. *SD* is in parentheses.

between these two AOIs make up 2 %, respectively 4% of all transitions during writing.

A so-called AOI sequence chart can provide further insights of how viewing behavior changed over the 30 minutes of task duration on the individual level (see Figure 2-7). In the case of participant 2-05, the AOI sequence chart shows that after an initial phase of the writing process, where he shortly looked at the reading text, the graphical input and the text box, the participant started to read the instructions. He then gazed at the graphical input, with some transitions to the text box. This was followed by a longer phase where he read the source text, repeatedly returning to the instructions as well as to the sources, i.e. the text and the graphic. After around six minutes (which is equivalent to 400.000 milliseconds in the AOI sequence chart), the participant mainly looked into the text box and at the graphical input. Beginning at minute 10, his views transited frequently between his own writing and the source text, interrupted only by revisiting the instructions. It's striking, that towards the end of the task duration, the participant basically dwelled in the text box, only sometimes looking at the remaining time at the very end of the 30 minutes.

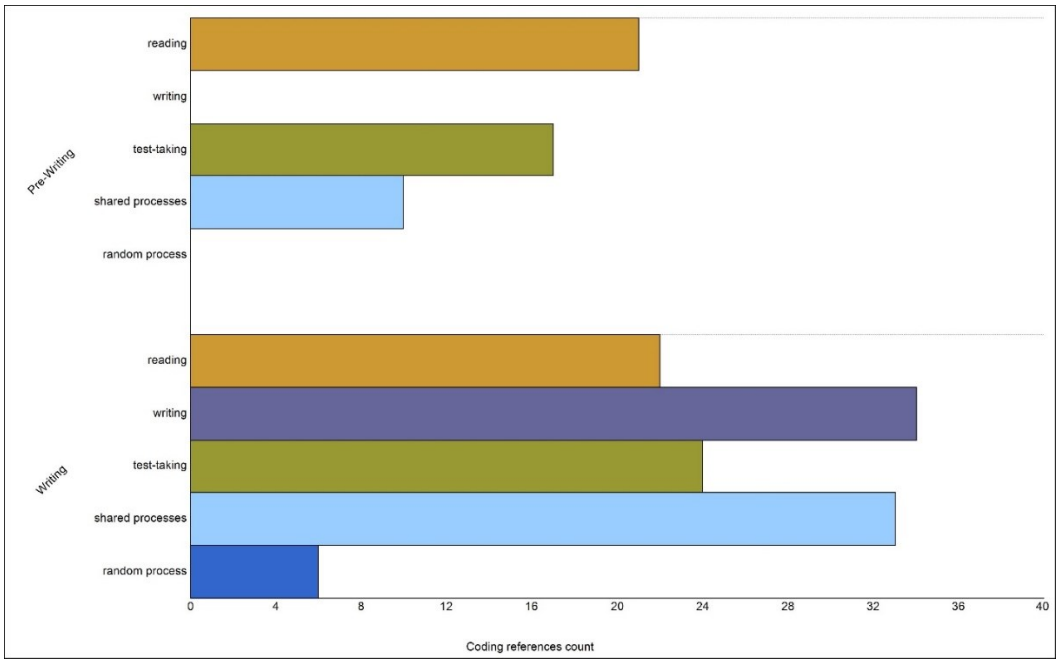
The data presented in an AOI sequence chart provides insights into the viewing behavior of participants over time, but it does not explain the cognitive processes involved. Further insights into the cognitive processes can be gained by looking at the stimulated recalls. A Matrix Coding Query in NVivo was run to explore the intersection of the nodes *Phase* and *Cognitive Processes*. The results of this analysis are presented in Figure 2-8.

Figure 2-7 *Example of an AOI sequence chart participant 2-05*



The figure illustrates that during pre-writing participants mainly engaged in reading and test-taking processes, i.e. processing the input material and reading the instructions before starting to write.

Figure 2-8 *Cognitive processes in relation to writing phase*



Test-taking strategies during pre-writing also included checking the remaining time, which most of the participants did after they finished reading the source text. Participant 1-09 for example looked at the timer when almost finished with reading the source text to decide how to proceed:

> Ja, ich möchte, ich möchte wissen, wie viel habe ich gelesen. Deshalb habe ich, deshalb will ich entscheiden, ob ich sollen weiter lesen, oder mit dem Text beginnt schreiben. [*Yes, I wanted to know how long I read. I wanted to decide, if I should go on reading, or if I should start to writing.*]

Participants were also engaging in shared processes, i.e. they were re-reading parts of the source text or the instructions again, looking for ways to e.g. phrase the introduction, like participant 2-06 explained:

> Ja, also, ich, das ist der Grund, warum ich zurück zum Text komme. Ich möchte eine Einleitung finden. [*Yes, that is the*

reason why I went back to the text. I wanted to find an introduction.]

During writing, of course, participants engaged in writing-related processes, but also to a great extend in shared processes and reading. The high amount of shared processes at this stage of the writing process is related to the engagement with the source material, reading involved processing the text and the graphical input for selecting relevant information. In their comments on test-taking strategies during writing, participants were to a great extent referring to time management. But whereas checking the remaining time during pre-writing was perceived positive in general as described above, this perception changed during writing, especially towards the end of task duration. Participant 1-06 described what he was thinking after one last look at the timer during the final minute of task duration:

> Hm, jetzt habe ich knappe Zeit. Jetzt, was soll ich schreiben? Ich habe nach (inc.) geguckt. Wie viele Wörter? Und, was jetzt soll ich machen in diese kurzen Zeit? Weiter Informationen von den genauen Text, von hier schreiben, oder was von Kopf, oder eine Schluss? Mache ich einen Abschluss, einen Schluss geschrieben. Also jetzt, soll ich nach die Grafik noch Vorteile, Nachteile nennen? Oder (inc.)? Was soll ich jetzt machen? Ok, die brauchen das nicht. Schreibe einfach einen Schluss. Ich habe keine Zeit, fertig. [*Hm, I only have little time left. What should I write now? I looked at (inc.) How many words? What should I do in this little time left? Write further information from the text, or own ideas, or a conclusion? I write a conclusion. Now, should I also add advantages and disadvantages after describing the graphical input? Or (inc.)? What should I do? Ok, they don't need it. Just write a conclusion. I have no time left, done.*]

This section showed that the cognitive processes during pre-writing were distinct from the ones participants employed during writing. While pre-writing was determined by reading processes and test-taking strategies, participants most frequently engaged in writing and shared processes during writing.

2.4.4 Effect of test-taker characteristics on cognitive processes

Two sub-questions focused on the effect of influencing variables like overall language competence (RQ 1d) and the typing skills of participants (RQ 1e).

To answer these research questions, descriptive statistics for both, the C-test results and the typing test, are reported. Comments from the stimulated recalls provide additional insights on these effects.

Language proficiency

As described in the overall research design (see section 1.5), the C-test was used to measure participants' overall language proficiency. With four texts and 20 gaps each, participants could score a maximum of 80 points in the C-test. It is apparent from Figure 2-9 that only two participants scored more than 65 points, the majority yielding scores between 30 and 50 points.

The test scores were also assigned to the CEFR levels A2 to C1. Looking at the according cut-scores (see Figure 2-10), one participant was placed at A2, nine at B1, five at B2 and two participants yielded „C1 and above" as a result.

Data was normally distributed for the two sets (Set 1 and Set 2) participants were randomly assigned to (Shapiro-Wilk, $p > .05$), with scores a little lower in Set 1 ($M = 47.13$, $SD = 9.49$) than in Set 2 ($M = 49.89$, $SD = 10.142$). There was homogeneity of variance (Levene's test, $p > .05$). An independent sample t-test revealed no statistically significant difference between Set 1 and Set 2, $t(15) = -.578$, $p = .572$.

Figure 2-9 *C-test results: Distribution of C-test scores*

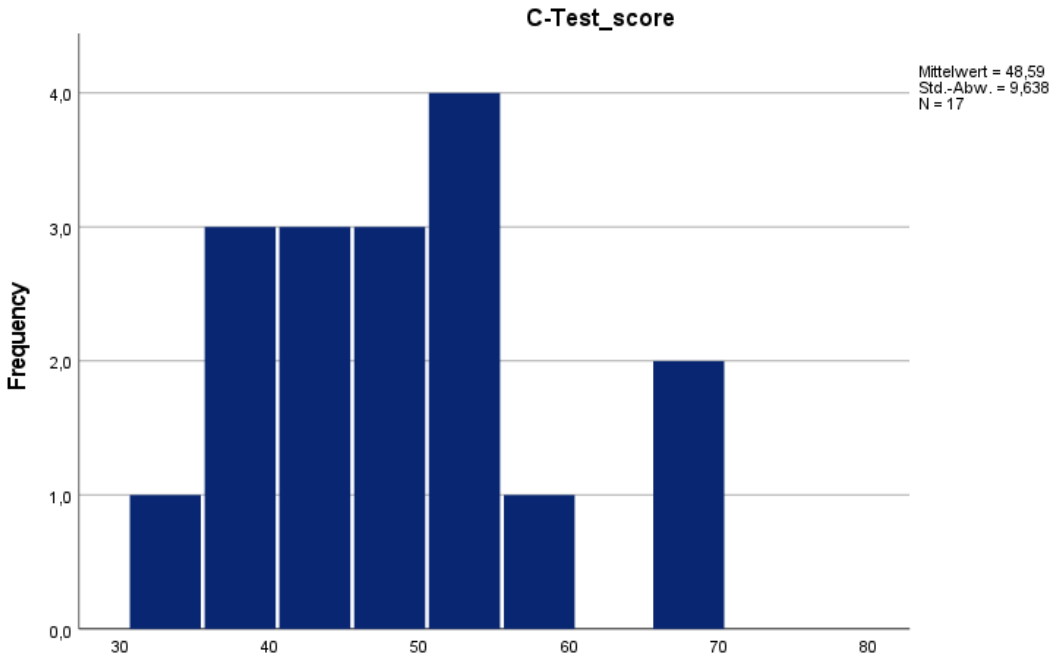
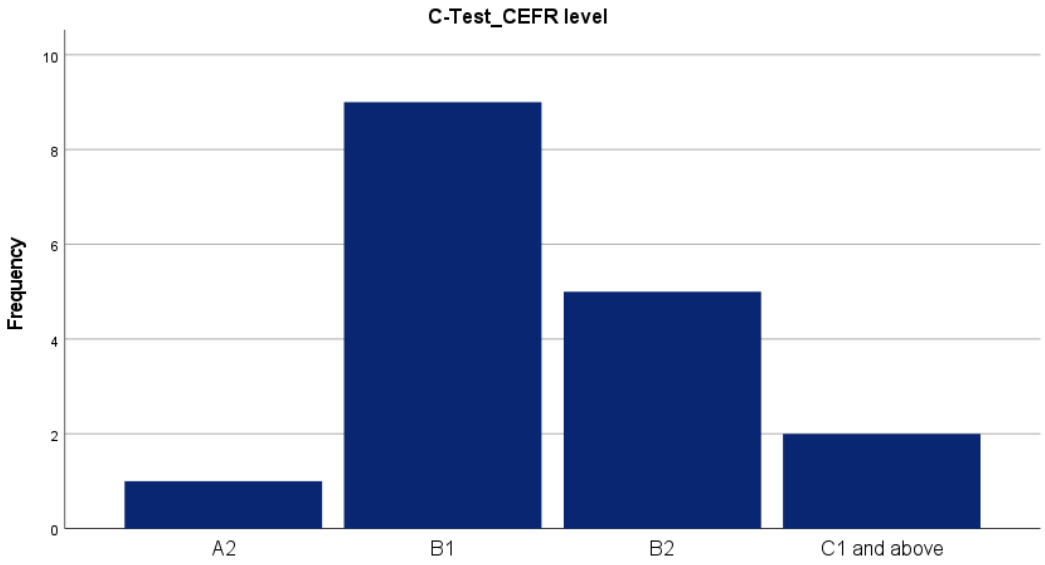


Figure 2-10 *C-test results: Distribution of CEFR levels*



In order to see how the overall language proficiency affected the processing of the task the sample was divided in two groups – one higher and one lower proficiency group. Out of all participants from

which both the eye-tracking data and the stimulated recalls were available, the four highest scoring participants in the C-test, i.e. those yielding B2 and above were placed in the higher proficiency group, whereas the lower proficiency group consisted of the four participants with the lowest C-test scores, yielding results of A2 and B1.

The analysis of how participants approached the task, i.e. the time they spent on pre-writing and writing, as well as the analysis of the eye-tracking metrics of dwell time, revisits and transitions revealed no statistically significant differences between lower and higher proficiency participants, except for the percentage of dwell time participants spent looking at the graphical input: There was a statistically significant difference for the dwell time in the graphical input between low (M_{Rank} = 6.50) and high (M_{Rank} = 2.50) proficiency learners, U = .000, Z = -2.309, p =.029, but with a negligible effect size (r = -.010).

In the stimulated recalls six participants commented on difficulties in processing the task due to language competence, interestingly not only participants who were in the lower proficiency group. On the one hand, the interviews revealed that language proficiency had an impact on the comprehension of the input material, especially the source text, as for example, participant 2-03 reported:

> Ok, die schwierigste ist der Text. Viele schwierige Wörter, und da viele Nebensätze, man muss denken, ok, was ist passiert, und viele fremde Adjektive. Und, äh, aus diesem Grund habe ich wenig verstanden, was, das der Text. [*Ok, the most difficult part is the text. Many difficult words, and many sub-clauses, one has to think, ok, what happened, and many unfamiliar adjectives. And that is the reason why I only understand so little about the text.*]

Almost all of these comments focused on challenges in comprehending single words; only some participants reported on problems that went beyond word level, like participant 2-06:

> Ja, diesen Abschnitt habe ich wirklich nicht so gut. Ich möchte die, eine kleine Zusamm, eine kleine Zusammenfassung, aber die Idee von diese Abschnitt, von diesem Abschnitt habe ich wirklich nicht so gut verstanden. [*Yes, this passage, I didn't really well. I wanted to make a short summary, but I didn't get the idea from this passage.*]

The participants not only reported language problems in relation to comprehending the sources, some participants also commented on this in relation to their own writing. Asked about the impression of the task, participant 2-06 explained that he was not satisfied with his own writing:

> Nein, vielleicht nur, mein Text besser schreiben, wie mit andere Wörter. Ja, ich weiß, ich habe nicht so gut, ja manchmal fehlten mir die Wörter, oder die Idee, was sollte ich sagen. Welche Wort sollte ich benutzen, auch wenn ich schon, was zu sagen hatte, wusste ich nicht, wie ich das sagen sollte. [*No, maybe that I want to write a better text, using other words. Yes, I know, sometimes I didn't have the words, or the ideas, to write what I should write. What word should I use, even when I had something to say, I didn't know how to write it.*]

Overall, the analysis of the eye-tracking data showed no differences in processing between low and high proficiency learners. But C-test results and the stimulated recalls revealed that for several participants the level of language competence was not sufficient to deal with the demands of the task. For some participants the lack of language hindered the full comprehension of the sources and had an impact on the summarization of relevant information in their own writing.

Typing skills and computer familiarity

Participants typing speed was around 20 WPM, with 211 characters being typed on average during the whole two minutes of

the test. Taking into account the numbers of mistyped words, the adjusted typing speed was a little bit lower with 15.50 WPM. To put these numbers into perspective: according to the feedback that comes together with the typing test results, an average typist reaches 36 WPM, an average touch typist, i.e. someone who does not need to look down at the keyboard, achieves 58 WPM.

Table 2.15 *Results typing test*

	WPM	**Number of chars in 2:00**	**Number of errors**	**Typing accuracy (%)**	**Adjusted typing speed (WPM)**
M (SD)	20.81 (6.76)	211.81 (68.59)	10.44 (13.85)	76.00 (26.00)	15.50 (7.82)
Mdn	20	204	4.50	87.50	14.50
Min.	10	109	.00	29	5
Max.	34	348	48	100	34

Note. N=16.

A Mann-Whitney-U-test was calculated to determine if there were differences in the adjusted typing speed between participants in Set 1 and Set 2. There was no statistically significant difference in adjusted typing speed between both groups, $U = 28.000$, $Z = -.421$, $p = .721$.

Nonetheless, there were considerable individual differences amongst the participants. Looking at the WPM and the total numbers of characters participants were able to type during the two minutes of the test, it also becomes evident, that their typing accuracy varied. While some participants made no mistakes at all, leading to an accuracy of 100%, other mistyped 48 words which resulted in a typing accuracy of only 29%. According to the overall typing speed feedback, 15 out of the 16 were slow typers, only one typed with an average typing speed.

The challenges in typing that were evident from the direct measure of the typing test were also confirmed by self-reported difficulties of some participants in the stimulated recalls. These difficulties in typing were related to problems with typing in general, but were also linked to the keyboard type, like participant 1-02 reported:

> Ähm, für mich, alles am Computer schreiben, mit Tippen ist etwas neu. Also deswegen find ich, war es schwierig, die rechte Buchstaben zu finden und wenn ich möchte Buchstaben groß machen, ich habe immer die falsche Tastatur gedrückt. Also das war schwierig für mich. Ich kann quasi gut schreiben, tippen an einer englischen Tastatur. [*For me, typing on the computer is somewhat new. That's why it was difficult to find the right letters, and when I wanted to write something in capital letters, I always pressed the wrong key. That was difficult. I can write well, type on an English keyboard.*]

Besides problems with typing, the retrospective interviews also revealed another aspect that impacted the processing of the task that was related to the medium of the computer. Participants said that the processing of the input text was different on the screen compared to paper-based texts, as participant 1-09 described:

> Wenn ich mit die Papiere lesen, lese, dann kann ich schreiben Notizen besser. Und vielleicht es gibt, ich weiß nicht, wenn ich Notizen ist Papier, kann ich besser verstehen als wenn ich in, na mit dem Computer tippen. Deshalb muss ich lesen nicht einmal, aber zweimal oder dreimals um die Text zu verstehen. Das ist ein Problem. [*When I read something on paper, I can take better notes. And maybe, I don't know, when I take notes on paper, I can better understand compared to when typing on the computer. That is why I not only had to read once, but twice or even three times to understand the text. That is a problem.*]

It seems that while reading on paper, readers not only build a mental but also a spatial representation of the text, i.e. they seem to remember where a certain piece of information in the text can be found. But processing a text on a screen seems to be different, as participant 2-07 explained while being asked why he highlighted information in the source text:

> Ne, eigentlich, normalerweise wenn ich einen Text schreibe, oder Zusammenfassung schreibe, dann markiere ich nichts. Aber ich dachte, ja vielleicht, das habe ich HIER gemacht, weil ich bin daran gewöhnt, dass ich auf einem PAPIER eine Aufgabe zu haben. Aber hier, vielleicht verliere ich, wo war die Information. Deshalb habe ich das markiert. Damit ich nicht zu viel Zeit verbringe, die Informationen zu finden. [*Normally, if I write a text or a summary, I don't highlight anything. But I thought, yes, maybe I did it HERE because I'm used to paper-based tasks. But here, I might not remember where the information is in the text. That is why I highlighted is. Not to waste time to find the information.*]

Besides the typing difficulties evident in the results of the typing speed test, the stimulated recalls not only backed these findings, they also revealed to some extent that processing tasks online differs from processing tasks on paper. This is in line with existing research on the effect of test delivery mode on construct equivalence of paper-based and computer-delivered tests (see section 1.3).

Washback effect

Even though not in the focus of the research questions for this strand, participants mentioned washback effects as another potential effect on their processing in the stimulated recalls. Except for one, all participants were enrolled in preparation classes for the paper-based TestDaF. But since the integrated writing task in the digital version was so different from their former writing experience, seven out of 14 participants reported on struggling with the task since they could not simply transfer the writing strategies they had acquired for the paper-based test, as participant 1-09 explained:

> Äh, es gibt meine ich, auch zwei Probleme, dass ich nicht daran, an diese Aufgabe gewöhnen. Deshalb musste ich auch, weil wenn ich vielleicht TestDaF, die TestDaF-Prüfung mache, habe ich schon erfahren die Struktur, und die ganze Redemittel schon gelernt haben. Deshalb kann ich sehr schnell schreiben. Aber mit diese Aufgabe, ich habe Probleme. [*I think there are two problems, I'm not used to this task. When I sit the TestDaF, I know the structure, and I have learnt useful phrases. That's why I can write fast. But I had problems with this task.*]

Since participants were not familiar with the task and the genre „summary", they had to (re-)read the instructions carefully to define the task, as reported earlier, and to match the task requirements with their former writing experience, like participant 1-05 described:

> Und für mich war ein bisschen schwer, weil wir, äh, unsere Aufgaben haben, und nicht zur Zusammenfassung. Und ich weiß nicht, zum Beispiel, wenn ich meine Heimatland beschreiben möchte, darf ich das machen, oder nicht? Und das passt dazu, oder nicht? Und in welche Abschnitt muss ich das legen, zum Beispiel am Anfang, oder am Ende? [*For me it was a little hard, because we have our tasks, but no summary. For example. I didn't know if I was allowed to write something about my home country. Does it fit here, or not? And where do I put this, at the beginning, or at the end?*]

In the paper-based writing component of the TestDaF, participants have to write a coherent, argumentative text on a given topic. The instructions require that they should describe the situation in their home country according to the topic. There is evidence that test takers internalize the structure of the task and memorize patterns in order to yield high scores on the test, but they often lack a real writing competence beyond this *teaching to the test* (Zimmermann, 2009). This may have happened here. Participant 1-05 had difficulties adopting to this new task type, and fell back on his/her prior writing experience.

2.4.5 Generalizability of cognitive processes across different test versions

The last research question (1e) examines possible task effects, i.e. whether the involved processes are generalizable across different versions of the task. To test for significant differences in the quantitative eye-tracking data between the two sets, non-parametric independent samples Mann-Whitney U tests were used to analyze the following data: (1) pre-writing and writing time, (2) dwell time, as

well as (3) revisits in the different AOIs, and (4) transitions. The results are displayed in Appendix F.

Test takers in both groups approached the task in a similar way, spending around 4 minutes on pre-writing and nearly 26 minutes on writing. There were differences between the two sets in the time they spent looking at the different AOIs, as well as in the frequency of revisits to the AOIs. Also the transitions between the AOIs were slightly different between Set 1 and Set 2. But all the differences were not statistically significant, except for the transitions from the timer to the source text: There was a statistically significant difference in median of transitions from the remaining time to the source text in Set 1 (*Mdn* = 1.48) and Set 2 (*Mdn* = 3.07), U = 64.000, Z = 2.694, p = .006 with a large effect size[30].

2.5 Discussion

The overall research aim of this strand was to investigate the cognitive processes of writers when working on the integrated writing task of the digital TestDaF. The analysis of eye-tracking and stimulated recall data revealed that during the whole writing process, participants used reading and writing processes, employed specific reading-writing processes and made use of test-taking strategies.

Results show that use of cognitive processes varied at different stages of the writing process, confirming findings from existing research (e.g Barkaoui, 2015; Michel et al., 2020; Plakans, 2008). Even though there were individual differences, participants approached the task in a similar way. At the beginning of task duration, i.e. during pre-writing, participants followed a rather

[30] According to Cohen's classification of effect sizes values of .5 and above show a large effect (Cohen, 1988, 1992).

linear process of reading the instruction and the source material, using mainly reading processes in order to comprehend the source material, and employing test-taking strategies to align their writing to the task demands. The actual writing phase appeared to be a more recursive process. Participants often paused for planning or revising, or they went back to the source material. While engaging with the sources at this stage of the writing process, participants employed reading processes to select and verify the relevant information for their summaries, or they used shared processes like "writerly reading" to mine the input material for language material they could use in their own writing. This use of the input material confirmed findings from previous research. For example, Wang (2018) describes this process in his study as a expeditious reading for lexical or content support. Even though revision took place during the whole writing process, test takers specifically focused on their own writing at the end of task duration for revision purposes.

It became evident, that reading plays an important role in the writing process. Reading is required for understanding the task instructions which helps writers conceptualize the task demands (see Wolfersberger, 2013). Reading is also involved in the comprehension of source material. As the viewing behavior of participants revealed, participants spent most of their pre-writing time reading the source text. During writing, they often revisited the source text and the graphical input, proving an engagement with the input material throughout the whole writing process.

In some cases, the comprehension of source material was hindered due to language problems as participants admitted in the stimulated recalls. The accompanying C-test confirmed this self-reported data, and showed that participants were mainly placed below the threshold level of B2. The excerpts from the retrospective interview revealed that when reading for comprehension,

participants were mainly engaging in bottom-up processes as described in Plakans (2009), Plakans et al. (2019), or Li (2014). However, the effect of language proficiency on the cognitive processes, e.g. indicated by Plakans (2008, 2009) could not be confirmed in this study.

An additional typing test was used to measure participants' keyboarding skills to examine possible effects on the cognitive processes. Results showed that participants lacked proficient typing skills. The stimulated recalls revealed that this was partly related to the unfamiliar keyboard layout which forced them to look down at the keyboard quite often. The effect of computer familiarity and keyboarding skills on the cognitive processes of writers in the context of the digital TestDaF therefore need to be further examined to back up these preliminary findings. If keyboarding skills are regarded as an integral part of the L2 academic writing construct (see section 1.3), international study applicants have to practice this relevant skill in preparation for university. This calls for writing on the computer to become an essential part of preparatory language classes. Regarding integrated writing, this holds also true for reading onscreen. In the stimulated recalls of this study, participants commented on the differences in processing text on paper and on the screen, confirming findings from previous research on the effect of the medium (see section 1.3). Language learners therefore need to develop effective strategies for reading online. Otherwise the lack of these strategies and the lack of sufficient keyboarding skills can be regarded as a threat to validity.

Overall, the analysis of cognitive processes in this study proved that the construct underlying the integrated writing in the digital TestDaF is a specific interaction of basic reading and writing processes that is affected by personal characteristics (e.g. language proficiency) and external factors like task demands (see Asención

Delaney, 2008). At certain points in the writing process, "reading and writing processes blend" (Spivey, 1990, p. 258) – a unique integration of skills in integrated writing.

The processes described in this study mostly confirmed findings from existing research in the EFL context and would allow for further linking to theoretical models like *discourse synthesis* (see Spivey & King, 1989), implying that the current understanding of integrated writing can be expanded to other languages than English.

2.5.1 Limitations

Of course this research study had its limitations. First of all, the sample size was relatively small with 19 participants. However, it should be considered that the combination of eye-tracking and stimulated recalls produced a large amount of quantitative and qualitative data, and that the analysis was quite complex, i.e. labor-intensive and time-consuming. Therefore, this specific research design can only be applied to samples with a limited number of participants.

The limitation of sample size was partly offset by the representativeness of the sample for the expected test-taker population of the digital TestDaF in relation to age, gender, regional distribution and study experience. In addition, two different versions of the integrated writing task were administered, so that results were generalizable across multiple task versions.

Another limiting factor was the relatively low overall language proficiency of the participants. As the C-test results showed, most of them were below the required threshold level of B2. This might have had an impact on the cognitive processes while working on the

integrated writing task, as well as on the verbalization of these processes in the stimulated recalls.

Participants were not familiar with task since data was collected during an early piloting stage of new test material for the digital TestDaF. At this point, the accessibility of test preparation material was limited. Participants had access to a model test, but no description of the test format or other information material was publicly available yet. This definitely affected their cognitive processes as became apparent in the stimulated recalls. Participants tried to apply their existing writing strategies when working on the integrated writing task. Partly this was also related to a washback from preparing for the paper-based version of the TestDaF.

There were also limitations related to the research methodology: Regarding eye-tracking, several factors should be considered when interpreting the results. Data was collected under real exam conditions, i.e. participants worked on the integrated writing in the original test environment. For this reason, it was not possible to build in a recalibration of gaze positions into the data collection "to ensure that accuracy is going to be maintained at that level throughout the experiment" (Conklin et al., 2018, p. 24). Head movements during the 30 minutes of task duration as well as frequent looks away from the screen to look down at the keyboard might have had an impact on the accuracy of gaze positions. In addition, the original test environment did not allow for enlarging the line spacing in the reading text. Therefore, the focus of the eye-tracking data was not on single words but rather on relatively large AOIs, which reduced to some extent the issue of accuracy.

A fundamental consideration in interpreting eye-tracking data is the way information is presented on the screen. For example, the "salience, size, and position of visual information on the screen affects the likelihood of fixation" (Godfroid & Hui, 2020, p. 280). In

the integrated writing task of the digital TestDaF the graphical input is prominently placed in the middle of the screen and the only element of the task that is presented in color (see Figure 1-1). This task layout might affect the processing of the different sources, but the effect could not be observed in the data of the current study.

In relation to the stimulated recalls, several limitations exist: Firstly, due to time constraints and for logistic reasons, the stimulated recalls had to take place immediately after the eye-tracking experiment. On the one hand, such a short period of time between the event and the retrospective interview is seen as an advantage since the accuracy of the recall is usually high (Gass & Mackey, 2017). On the other hand, participants did not receive comprehensive training, and the immediate recall also did not allow for a careful review of the gaze replay videos before the stimulated recall sessions by the researcher. Instead, parts of the writing process that were used as a stimulus in the retrospective interviews had already been pre-selected during the actual writing process. In contrast to unstructured, participant-led recall sessions used in other studies (e.g. Michel et al., 2020; Wang, 2018), the semi-structured retrospective interviews reduced the cognitive burden on participants and helped in limiting the amount of data (see also (Asención Delaney, 2008; Yu et al., 2017). While the semi-structured approach might have also pose a risk to overlook all processes involved in integrated writing, participants in the current study were asked if they could remember other relevant moments in their writing process which were not captured in the interviews at the end of the recall sessions.

Secondly, the stimulated recalls had to be conducted in German, the heterogeneous language backgrounds did not allow for conducting the retrospective interviews in the L1 of the participants. This might have impacted the verbalization of cognitive processes,

especially since many participants were not highly proficient in German as the C-test results revealed.

Even though the research revealed some limitations in interpreting the results, the current study still provided valuable insights in the cognitive processes of test takers completing the integrated writing task of the digital TestDaF.

Process can not be inferred from product any more than a pig can be inferred from a sausage.
(Murray, 1982, p. 18)

3 A PRODUCT-ORIENTED APPROACH TO VALIDATION

The previous chapter approached the construct underlying the TestDaF integrated writing task from a process-perspective. This chapter takes a closer look at the written performances. By analyzing the written summaries of participants who took part in the process study, and following the above outlined argument-based approach to validation (see chapter 1.4), the product-oriented strand of this study tries to provide evidence for the assumption that the (successful) processing and transformation of the input material is evident in the written products.

The first section of the chapter outlines the effect of different factors that impact the quality of written responses in integrated writing assessment. An overview of relevant research is provided in section 3.2. The subsequent section (3.3) describes the methodology of the current study, findings are reported in section 3.4. The chapter closes with a discussion of results.

3.1 Variables accounting for the quality of written performances

Different factors do not only affect the writing process but also have an impact on the written output. This section therefore takes a closer look at two of the factors that impact the written output with the focus on integrated writing assessment: test-taker characteristics and task variables. Test-taker characteristics include e.g. background knowledge or language competence, task variables comprise aspects like the expected genre or the type and complexity of the input material.

Test-taker characteristics

When it comes to test-taker characteristics, one of the most influential factors that affects the written output is the overall language proficiency of the writer – but according to Asención Delaney rather „as an additive to other variables […] than as a single causative factor" (Asención Delaney, 2008, p. 142).

To explain the role of L2 proficiency, many studies have therefore looked at the differences between L1 and L2 writers (Keck, 2006; Shi, 2004), or have compared low-level and advanced writers within L1 or L2 (e.g. Chin, 2009; Plakans & Gebril, 2013; Weigle & Parker, 2012).

In integrated writing assessment, L2 proficiency affects the writing in different ways. Text comprehension is crucial, and "if writers are not competent readers, if they oversimplify or misunderstood the source text, their own texts that interpret or summarize those source texts are likely to suffer" (Hayes, 1996: 18). In addition, L2 writers may encounter problems in reformulating the ideas from the sources. This becomes evident in the written products, particularly when looking at integration style, i.e. how the information from the input material is incorporated in the written output. For example, L2 writers tend to do more direct copying instead of paraphrasing when transforming the language of the input material. (Grabe & Zhang, 2013; Keck, 2006; Shi, 2004).

Related to the whole issue of integration style is the question, how appropriate the verbalization of information is. For instance, in the context of term papers of international students in Germany, Stezano Cotelo (2003) showed that reformulating of ideas was a problem for non-native speakers of German since they were lacking the linguistic resources. The attempt to transform the language of the source text led to imprecise expression of ideas (due to morpho-

syntactic or lexical errors) and to formulations that had similarities with language produced in the spoken domain. Therefore, students tended to use citation and direct copying from sources as a strategy to overcome these hurdles (Stezano Cotelo, 2003, p. 111).

To see inappropriate source use – in the sense of plagiarism – as a strategy of writers to deal with language problems rather than academic dishonesty fits well into recent views on that issue. In this view, researchers have argued that copying from sources, or *patchwriting*, i.e. "copying from a source text and then deleting some words, altering grammatical structures, or plugging in one-for-one synonym substitutes" (Howard, 1993, p. 233; cited after Wette, 2017), can be regarded as a stage in skill development to become a proficient writer (see Pecorari and Petrić (2014) for an extensive discussion on plagiarism on L2 writing).

Another factor that also has an impact on how writers use the sources and how they present the information in their own writing is background knowledge. If writers are familiar with the topic, it might be easier for them to process the sources, either because they have content knowledge, or they know specific vocabulary associated with the topic which facilitates source text comprehension. For example, in the context of a reading comprehension test for German as a foreign language, Krekeler (2006) found a strong effect of background knowledge on the scores, as students with background knowledge outperformed the ones without.

Finally, in addition to background knowledge on the topic, writers also need to develop an understanding of the task requirements, also known as *task representation* in order to work successfully on a task (Plakans, 2010; Wolfersberger, 2013).

Task variables

Apart from characteristics of the writer, the task itself affects the written outcome in multiple ways, particularly through characteristics of the source material and task type. Looking at the sources of different integrated writing tasks in use, one can see that input material differs with regard to discourse type (i.e. if a text is mainly narrative, expository or argumentative), text length, linguistic complexity and organizational features which all contribute to the readability of a text. With regard to summarization tasks, Yu (2009) looked at how certain qualities of source texts contribute to their summarizability, and therefore have an impact on summary writing in the L1 (Chinese) and L2 (English). Although the input material was of similar length and readability, discourse features like lexical diversity and macro-organization differed across texts. Even though the analysis of the performance data showed a significant effect of source text on students' summaries, post-summarization questionnaires and interviews could not reveal one single factor accounting for this effect. Yu therefore came to the conclusion that source text characteristics and interpersonal characteristics of the summarizers interact in various ways.

The written output is also depending on the availability of the sources during task completion and the expected genre that test takers have to write. Regarding the availability, Yu (2009) reported on studies that discussed deeper learning effects in text-absent in contrast to more direct copying in text-present summarization (ibid., p. 118). Cho, Rijmen, and Novak (2013) also voiced that the responses on the TOEFL iBT integrated writing task might have been influenced by the fact that the reading passage was presented during the whole time of task completion, while test takers listened to the lecture only once.

Asención Delaney (2008) showed in her study on two different integrated writing tasks – a summary task and a response essay – that scores on both task types have only a weak correlation. She therefore concluded that these different task types measure different dimensions of reading-to-write ability. In other words, the expected genre has an effect on the written output. The reason behind this may be that the response essay is more challenging for low-level learners since it involves more critical thinking skills, whereas summary writing as a descriptive rather than an argumentative genre is considered to be easier in terms of cognitive processing and writing ability.

To sum up, there are many factors that impact the written output of integrated writing tasks. Regardless of the complex interplay of test-taker characteristics and task variables, many studies have looked into the written products to see what kind of discourse features are evident, and whether or not these account for differences in ability levels.

3.2 Product analysis in integrated writing assessment

There are several strands of research examining the quality of written performances on integrated writing tasks. Many studies have looked into the relation of discourse features across different performance levels and test scores. Another large body of research has examined the extent to which writers have modified the language of the input material for their own writing to avoid plagiarism. Other studies also investigated the extent to which writers made use of the source material to include relevant ideas in their performances. The following section will briefly review the existing literature, with a focus on studies related to integration style and source use.

Discourse features

To shed light on the effect that linguistic features measured in integrated writing performances have on the scores, studies have examined measures of complexity (e.g. lexical diversity or syntactical complexity), accuracy or fluency across different levels of proficiency (Gebril & Plakans, 2009; Gebril & Plakans, 2013; Plakans & Gebril, 2017; Plakans, Gebril, & Bilki, 2019), sometimes in comparison with features measured in performances on independent writing tasks (Cumming et al., 2005). These studies have been carried out in EFL contexts, mostly using writing performances from the reading-listening integrated writing task from the TOEFL iBT.

Results from these studies mainly support the hypothesis that higher scores in integrated writing assessment are related to more complexity, accuracy or fluency in writing, even though findings are inconclusive. For example, Cumming et al. (2005) found that lexical diversity differed significantly across score levels, but for syntactic complexity a significant difference was only found in words per T-unit, but not in clauses per T-unit. In their study on organizational pattern, coherence and cohesion, Plakans and Gebril (2017) noticed higher scores with increasing quality of organization and coherence, while statistical differences across score levels could not be observed for cohesion markers. Discourse features were also identified as significantly different across performance levels by Gebril and Plakans (2013), although with some limitations. While fluency increased with increasing score level, measures like lexical sophistication or syntactic complexity did not. In addition, grammatical accuracy and source-use features differentiated well between low-level performances and mid- and high-level performances, but yielded no significant differences across mid- and high-level performances. Gebril and Plakans concluded that certain discourse features play a more prominent role in lower-level

performances, while other textual features become more critical at higher levels.

Comparing lexical diversity in independent and summary writing of L2 learners and native speaker experts, Yu (2013b) found that lexical diversity is a stronger predictor for independent writing than for summary writing. He concluded that lexical diversity is different in integrated tasks, because the vocabulary used is rather reproductive than productive knowledge.

Integration style

By analyzing texts of L1 and L2 speakers of English, Shi (2004) compared textual borrowing in summary and essay writing. The written performances of the students were coded for instances and the extent of textual borrowing, depending on whether they exactly copied strings of words from the source material, whether they modified the language of the input material slightly on the word level, or whether they paraphrased the original text using syntactical reformulations. Coding also took into account the acknowledgment of sources, i.e. if participants referenced the source text or not. As a result, a two-way analysis of variance (ANOVA) not only identified task effects and influence of the L1 but also an interaction between these two variables. The analysis revealed that more textual borrowing was found in the summary writing, and that English as a second language (ESL) students borrowed significantly more words from the sources than native speakers of English.

Weigle and Parker (2012) used an adapted scheme from Shi (2004) to investigate source text borrowing in an integrated ESL reading-writing assessment. Results showed that textual borrowing was limited to short strings of words from the reading texts or rephrasing the prompt, regardless of the educational level of

participants, i.e. undergraduate and graduate students. Topic effects became only apparent after close examination of the borrowed phrases. There were no statistically significant differences across different levels of proficiency, but students with lower scores "tended to quote more extensive excerpts from the passages when they did quote from the sources" (ibid., p. 128). As Weigle and Parker pointed out, this is contrary to the findings of Cumming et al. (2005), but maybe related to the fact that "different task types appear to elicit different borrowing strategies" (Weigle & Parker, 2012, p. 129; on the effect of task on paraphrasing see also Shi, 2004).

Taking into account different levels of paraphrase, Keck (2006) adopted the construct of *attempted paraphrase* which she defined "as an instance in which a writer selects a specific excerpt of a source text and makes at least one attempt to change the language of the selected excerpt" (Keck, 2006, p. 263). According to Keck, the development of a taxonomy of paraphrase types based on the construct of *attempted paraphrase* allows to distinguish between different levels of paraphrasing in source-based writing performances of L1 and L2 writers. She used this taxonomy in her study to compare summary writing performances of L1 and L2 university students. The results showed there were no significant differences between L1 and L2 writers in using attempted paraphrases in their summaries, but the two groups significantly differed in the use of various paraphrase types. L2 writers used more *Near Copy* paraphrases than L1 writers, i.e. paraphrases "composed primarily of long copied strings taken from the original excerpt" (Keck, 2006, p. 268). This finding is consistent with Shi (2004) who showed that summaries of L2 participants contained more and longer strings of words taken from the original text than L1 writers.

Another study that looked into language use in integrated writing assessment is that by Ohkubo (2009) who examined the

acknowledging and reformulations of the source material in integrated writing performances at different levels of proficiency. An in-depth analysis of six test takers' responses to a TOEFL iBT practice test showed that higher-scoring participants were able to attribute the sources in their texts, using different kinds of reporting verbs, while less successful test takers did not acknowledge the sources. The extent of reformulations was dependent on the source type: There was no evidence of direct copying from the lecture in any of the test takers' essays, even higher-scoring students only made some minor lexical or syntactical changes to transform the language of the oral input. The reformulation of the reading text unexpectedly revealed that lower-level participants were more likely to reproduce the ideas from the input material with their own words, while texts of successful participants again were closely based on the phrasing of the reading texts. Ohkubo explained this surprising finding with the fact that the formulations based on the input material used by higher-scoring students were more correct compared to the imprecise language of participants who used their own words – and hence were assigned with higher scores by the raters.

Source text use

Compared to the number of studies that have focused on linguistic features and on integration style, research that has solely looked into the quality of integrated writing performances with respect to content is scarce. Some studies have looked into source use as a follow-up in investigating writers' processes (Plakans, 2009a; Plakans & Gebril, 2012), with inconclusive results. Based on the frequencies in the use of discourse synthesis processes, Plakans (2009a) analyzed the frequencies of idea units with source use in the written performances of her participants. She could observe that the writers who used more discourse synthesis processes also had a

higher percentage of idea units using the source texts compared to those participants who used fewer discourse synthesis processes. The relation of process and product data was confirmed by Plakans and Gebril (2012), who also looked into the effect of source use on test scores. Their analysis did not reveal a clear pattern of source use across score levels though, pointing out "the difficulty in interpreting the use of sources in responses, namely that more does not necessarily equal better" (ibid., p. 30).

Source use has also been investigated in conjunction with analyzing linguistic features of integrated writing performances. For example, Plakans and Gebril (2017) looked into summarization patterns of TOEFL iBT responses. They reported that low-level essays were not so balanced in summarizing from two sources, i.e. the listening and the reading input. These essays contained more information from the reading than from the listening, while higher scored essays were centered around information from the listening passage. Plakans and Gebril (2017) concluded that source comprehension plays a major role for source use in integrated writing performances, but they claim that the variance in low- and high-level writers may also have been affected by different task representations.

Chin (2009) combined content-related and linguistic criteria to compare summary writing performances of high-intermediate and advanced-level Taiwanese EFL students. With respect to content-related aspects, the results showed that advanced writers tended to write shorter, but more accurate, concise and coherent summaries. Less skilled writers included fewer main ideas, but more unimportant ideas – a finding that according to Chin corresponded with earlier research (e.g. Kintsch, 1990) which showed that less proficient summary writers can identify the overall information, but fail to differentiate finer levels of importance and get distracted by

"seductive details" (Garner, Gillingham, & White, 1989). Chin also reported significant differences between both groups in terms of integration style. The advanced-level learners demonstrated better paraphrasing and integration skills, while the high-intermediate level participants copied more verbatim from the source text. With regard to language control, results from Chin's study suggest that a lack of lexical and grammatical control can affect the summary writing performance and lead to lower scores.

A combination of content- and paraphrase-coding was used by Wette (2017) who looked into different aspects of source text use in undergraduate EFL students' texts. She coded written assignments for the accuracy of the content (*accurate* vs. *inaccurate*) and for the extent of paraphrasing students used (from *no copying* over *some copying*, i.e. patchwriting, to *extensive copying*, i.e. plagiarism). With respect to paraphrase quality, her analysis revealed that to a great extent students used their own words to reproduce the information from the sources, while patchwriting and extensive copying were found to a lesser extent. The accuracy of reproduced information was very high, regardless of the type of paraphrasing.

In their study on the integrated writing task of the TOEFL iBT, Plakans and Gebril (2013) explored the integration of relevant information from the reading and the listening input by determining an *importance score* for each response. To do so, they rated the importance of every T-unit in the source texts, ranging from 4 ("very important", i.e. key idea) to 1 ("not important"). In a second step, they assigned this value to the corresponding T-units in the test takers' performances and established an importance score by totaling the essay T-unit scores. Essays were also analyzed for the origin of information, i.e. whether the information was taken from the reading or the listening source text, and for integration style. Results of their analysis revealed that source use and score are

related: High-scoring writers included more relevant information from the source material and made more use of the listening text by also using more paraphrasing, while lower-level writers depended more on the reading text and copied verbatim.

Cho and Choi (2018) explored the effect of audience awareness on three textual aspects: context statement, source attribution and content. They analyzed performances across different proficiency levels on two integrated reading-listening summarization tasks, one in which writers received specific information about for whom they were summarizing, one without specification of the audience. Results revealed that the quality of the included information was not affected by audience specification, but whether or not participants wrote for an unknown or a specified reader had an impact on the effectiveness of the context statement and on source attribution. All three aspects significantly differed across score levels within the writing condition with specified audience, i.e. participants with higher scores used more effective context statements, attributed the sources effectively and included more accurate information in their texts.

Overall, the existing body of research in the field shows that differences in the quality of integrated writing performances can best be explained by the level of test takers' language proficiency, but there is also an interrelation with task characteristics that accounts for variances in the written responses.

3.3 Methodology

3.3.1 Research aims and questions

This strand takes a closer look at the written performances of test takers who worked on the integrated writing task of the digital

TestDaF under exam conditions. The texts were analyzed in relation to content and to integration style in order to answer the following research question:

RQ 1: How do test takers process the information from the input material in the integrated writing task of the digital TestDaF? And how do they transform the language material from the two sources linguistically?

For further insights, the following sub-questions were formulated:

RQ 1a: To what extent are the written performances depending on test takers' characteristics? How important are background knowledge or task representation?

RQ 1b: Are there any differences in processing and transforming the sources between high- and low-level learners?

RQ 1c: Are the written responses affected by the comprehension of the source material?

RQ 1d: Can the results be generalized across task versions?

In a final step, this chapter explores whether product data can be linked to cognitive processes, by answering the following research question:

RQ 2: Are the inclusion of relevant information and the transformation of the input material in the written performances related to the viewing behavior of the participants?

3.3.2 Participants

The sample was identical with the one used for the process-oriented study in Chapter 2. For a detailed description of the sample turn to section 2.3.4.

3.3.3 Instruments

Besides working on the integrated writing task, participants also completed a C-test (see 1.5.2) to measure their overall language proficiency.

During the retrospective interviews that were used to investigate cognitive processes (see Chapter 2), participants were also asked (a) to comment on the task, and (b) to recall and summarize the topic and relevant information from the source text and the graphical input of the task they had just worked on (see Appendix B). The segments of the stimulated recalls dealing with these specific question were used in this strand for investigating (a) the effect of task characteristics, and (b) the comprehension of the sources independently from the ability to summarize the relevant information in written form.

3.3.4 Data collection

The written performances and the stimulated recalls were collected during the eye-tracking experiment described in section 2.3.6. The data set used in this strand consisted of 19 written performances and excerpts from 14 recordings and transcripts of the stimulated recalls (see Table 3.1).

Table 3.1 *Overview data set Strand* 2

Data collection	Data size used for analysis
completing the integrated writing task under test conditions (during eye-tracking experiment)	19 samples
excerpts from the retrospective interviews	14 recordings & transcripts

Performances of all 19 participants were included in the analysis, even though for some of the participants' data was excluded from the analysis of the eye-tracking experiment due to technical problems.

3.3.5 Data analysis

3.3.5.1 Written performances

Participants' written responses were analyzed with respect to content and integration style. These two aspects were considered important because they are also covered in the according rating scale.[31] The analysis should provide answers to the overall research question (RQ1), i.e. to explore how test takers process the information from the input material and how they transform the language material from the two sources linguistically.

Content

The content coding of the written performances partly adapted the coding scheme from Cho and Choi (2018), in addition some codes were derived from the rating scale of the digital TestDaF. The coding scheme is displayed in Figure 3-1. It included the following dimensions:

(1) *Context statement*, i.e. if participants made use of an introductory statement, and if so, how effective it was. If a text had a context statement, the sentence – or sometimes sentences – were coded with "effective" or "ineffective", depending if the introduction

[31] The rating scale currently in use for the digitale TestDaF is not published but is available upon request for interested researchers. Important aspects considered in rating the written performances are publicly available on the TestDaF website (https://www.testdaf.de/de/teilnehmende/der-digitale-testdaf/auswertung-des-digitalen-testdaf/; retrieved 09.10.2021)

provided enough background information on the topic. If a text started directly with the summarization of information without introducing the topic, "missing" was assigned to the whole response.

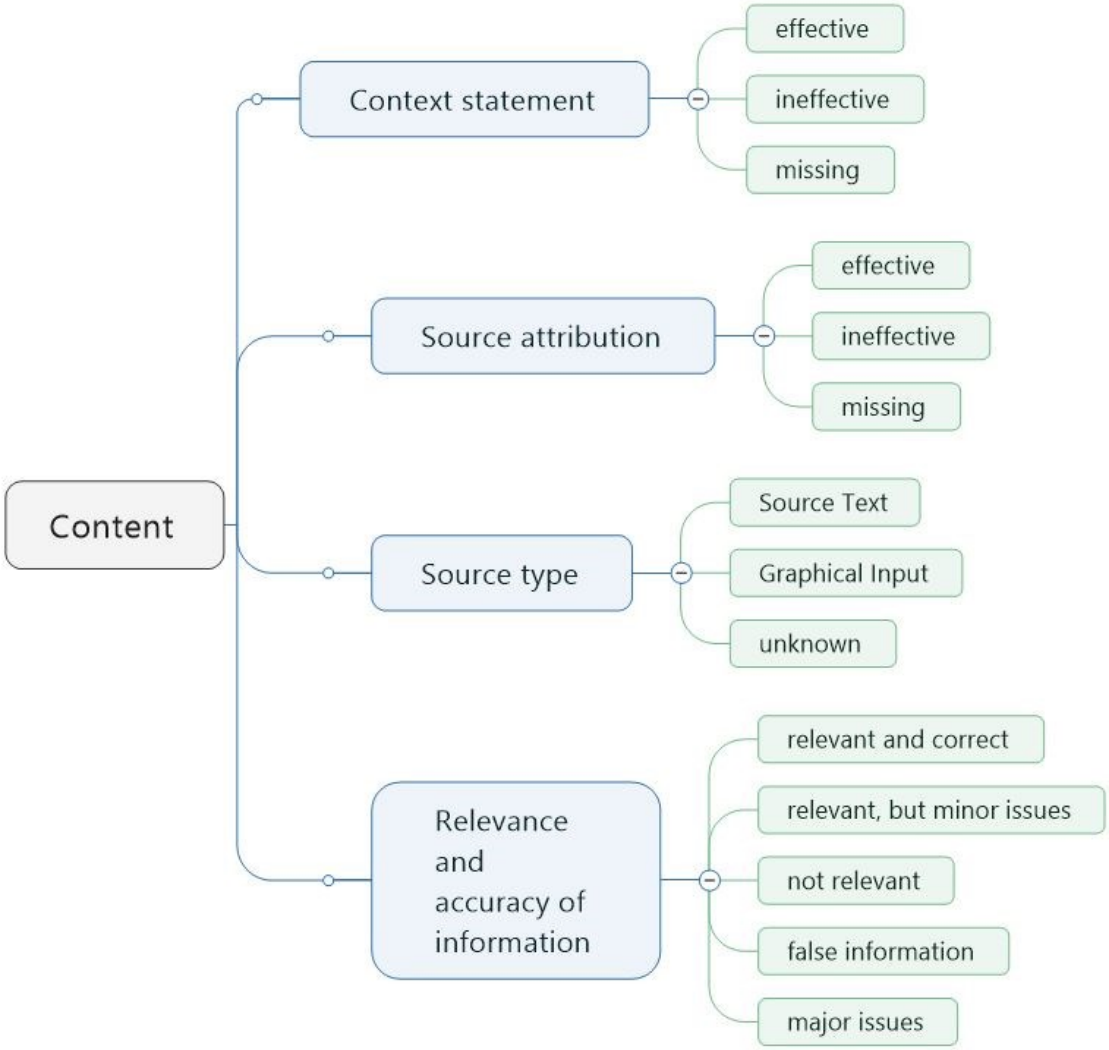
(2) *Source type*, i.e. if the information was taken from the source text or the graphical input. This coding was applied to each sentence of a written response outside the context statement. Since some test takers also included additional information in their texts, the category "unknown" was added, and applied to sentences in which information was not based on the sources.

(3) *Relevance and accuracy of information*, i.e. how relevant the information was with respect to the given question, and how precisely it was presented. The relevance of information was determined by using calibration material for TestDaF raters. This calibration material included benchmark performances and a description of items that are expected to be included in test takers' summaries. The material was established by a group consisting of experienced raters and internal experts at the TestDaF-Institut, using performances from piloting the specific tasks as benchmarks.

"Relevant and correct" was assigned when the information was important to answer the given question and was stated precisely. If there were minor imprecisions, e.g. an inaccurate number from the graphical input without falsifying the statement, the sentence was coded as "relevant, but minor issues". Parts of the performances that included information which did not answer the given question were coded as "not relevant", whereas "false" was assigned to information that was stated not correctly from the sources. When the information was not comprehensible due to language problems, it was coded as "major problems". As well as the source type, the coding of relevance and accuracy of

information was applied to each sentence of a written response outside the context statement.

Figure 3-1 *Coding scheme for content analysis*



The written performances were coded for *context statement* and *source attribution* to examine if participants were able to construct meaning from the task demands. Since the expected summary is intended to be a section within a chapter of a written assignment at university (see 1.2), an introductory statement as well as the

acknowledgment of the sources would be expected to guide the potential reader.

The coding of *source type* and *relevance and accuracy of information* was intended to provide insights into the processing of the two different input materials: Were test takers able to identify relevant information in relation to the given question? And how successful was the reproduction of this information? The decision to code *relevance* and *accuracy* in conjunction was based on the structure and phrasing of the rating scale for the integrated writing task in which these two aspects are combined in the content-related descriptor. For example, the content-related descriptor on level 4 describes a performance at this level as follows:

> enthält die meisten für die Aufgabe relevanten Informationen aus beiden Quellen, die weitgehend korrekt und strukturiert zusammengefasst werden [*contains the most relevant information from both sources; information is predominantly summarized correctly and structured*]

All coding was done using NVivo 12. Four out of the 19 performances (21% of the total sample) were randomly selected and double coded. The coding comparison in NVivo yielded the following results for inter-coder-agreement (ICR) at the primary level codes:

Table 3.2 *ICR for primary codes across double coded performances*

Primary Code	Agreement (%)	Cohen's Kappa
context statement	98.02	.96
source attribution	99.93	1.00
source type	98.76	.33
relevance and accuracy of information	98.76	.33

Inter-coder agreement (ICR) in terms of percentage agreement was very high with over 98% for all primary codes. Cohen's Kappa

coefficients revealed excellent agreement for "context statement" and "source attribution", while coding reliability for "source type" and "relevance and accuracy of information" was poor. Given the issues around the automated calculation of Cohen's Kappa in NVivo (see section 2.3.7 for a detailed discussion) and the high percentage of agreement, it was decided that the ICR for the coding of content was satisfying.[32]

Integration style

The written summaries were also analyzed regarding the extent to which participants copied language material from the reading text and the graphical input.

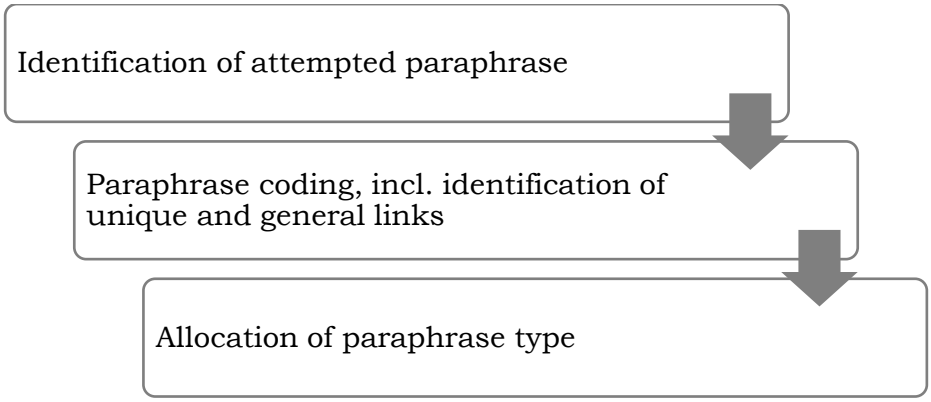
The coding for integration style in this study adopted the construct of *attempted paraphrase* by Keck (2006). She defined attempted paraphrase as "an instance in which a writer selects a specific excerpt of a source text and makes at least one attempt to change the language of the selected excerpt" (ibid., p. 263). According to Keck, this construct can be applied to examine different paraphrasing strategies, ranging from long strings of words copied from the source (as long as one change was made to the original source) to substantial changes in lexis or grammar with no copied strings at all. The analysis followed the procedure introduced by Keck, as displayed in Figure 3-2):

In a first step, the attempted paraphrases were identified manually by checking each sentence of the participants' responses against the source text, the title and the legend of the graphical input, as well as the instructions. Sentences that were not based on

[32] Cho and Choi (2018) also reported low Kappa values, especially for their content coding. They explained this with the high number of categories for content (N=11) that coders could apply.

the input or were simply typical formulaic sequences for structuring the responses like "Bevor zu diesem Thema Stellung genommen wird, werden einige Fakten anhand einer Grafik verdeutlicht. / *Prior to discussing the topic, some facts will be illustrated by using information from a graphical input.*" were not coded as attempted paraphrase.

Figure 3-2 *Stages of data analysis*



In a second step, all attempted paraphrases were coded based on the following characteristics: (1) number of words, (2) if a reporting phrase was used or not, and (3) the total number of *unique links* and *general links*. To identify *unique links* and *general links*, this study used Keck's definition who defined *unique links* as "individual lexical words (i.e., nouns, verbs, adjectives, or adverbs), or exactly copied strings of words used in the paraphrase that (a) also occurred in the original excerpt but, (b) occurred in no other place in the original text" (Keck, 2006, p. 266). But unlike in Keck's study, the unique and general links were identified manually and not by a computer program. The following example illustrates the construct of unique links[33]:

[33] Due to confidentiality of the test material, the following examples for the analysis will only be taken from test-takers' responses to the integrated writing task of Set 2 which is used as a model test within the "Training digitaler TestDaF B2/C1", an online German course offered by Deutsch-Uni Online (DUO).

Example 1 *Original*

> Gen-Mais bringt demnach weltweit, je nach Standort, durchschnittlich zwischen 5,6 und 24,5 Prozent höhere Erträge als konventionell produzierte Varianten.

Attempted Paraphrase

> Gen-Mais beträgt **durchschnittlich zwischen 5,6 und 24,5 Prozent als** traditionelle **Varianten**.

In this example, the words in bold are unique links related to the specific sentence in the source text. Even though "Gen-Mais / *genetically modified maize*" is also used in the original sentence, it was not identified as a unique link. Since the topic of the task is about genetically modified products, phrases like "Gen-Mais" or "gentechnisch verändert / *genetically modified*" are used frequently in the input material. These lexical words or phrases that occurred several times and could not be linked to specific excerpts in the input material were therefore identified as g*eneral link*s which "are more likely to be words associated with the important main ideas of the source text" (Keck, 2006, p. 267). In the following example, general links are underscored:

Example 2 *Original*

> Nur Europa ist eine Ausnahme, denn hier sind die Verbraucher gegenüber gentechnisch veränderten Lebensmitteln sehr kritisch.

Attempted Paraphrase

> Die Haupaussage des Textes ist folgende: Die Konsumenten sind doch gegen gentechnish veränderten Pflanzen.

The attempted paraphrase contains the words "gentechnisch / *genetically*" and "verändert / *modified*", which also can be found in the original source. However, since the term "gentechnisch veränderte Pflanzen / genetically modified plants" is a fixed term

that can be found multiple times throughout the source text, it was identified as a general link. If reporting phrases were used, like in this example ("Die Hauptaussage des Textes ist folgende / *The main message of the text is the following*"), they were not included in the total paraphrase word count.

In a final step, a paraphrase type was allocated to each attempted paraphrase based on Keck's taxonomy (see Table 3.3).

Table 3.3 *Taxonomy of paraphrase types*

Paraphrase type	Linguistic criteria
Near copy	50% or more words contained within unique links
Minimal revision	20-49% word contained within unique links
Moderate revision	1-19% words contained within unique links
Substantial revision	No unique links

Examples 3 and 4 will be used in order to demonstrate how the linguistic criteria were calculated for the attempted paraphrases:

Example 3 *Original*

> Gen-Mais bringt demnach weltweit, je nach Standort, durchschnittlich zwischen 5,6 und 24,5 Prozent höhere Erträge als konventionell produzierte Varianten.

Attempted Paraphrase

> Gen-Mais beträgt **durchschnittlich zwischen 5,6 und 24,5 Prozent als** traditionelle **Varianten**.

The attempted paraphrase in this example contains 11 words in total. There is one general link (underscored) consisting of one word ("Gen-Mais"), and two unique links (in bold) with a total of eight words ("durchschnittlich zwischen 5,6 und 24,5 Prozent als" and "Varianten"). In total, the paraphrase contains 81,82% (9/11) words that are copied from the input material. The percentage of the attempted paraphrase made up of words within general links is

9,10% (1/11), and the percentage of the attempted paraphrase made up of words within unique links is 72,73% (8/11). According to Keck's taxonomy, this attempted paraphrase was therefore classified as a *near copy*.

In example 4, the reporting phrase ("Die Hauptaussage des Textes ist folgende") was not included in the total word count, therefore the attempted paraphrase in this example contains eight words.

Example 4 *Original*

> Nur Europa ist eine Ausnahme, denn hier sind die Verbraucher gegenüber gentechnisch veränderten Lebensmitteln sehr kritisch.

Attempted Paraphrase

> Die Haupaussage des Textes ist folgende: Die Konsumenten sind doch gegen gentechnish veränderten Pflanzen.

There is one general link, consisting of two words, and no unique links. So the percentage of the attempted paraphrase made up of words within general links is 25 % (2/8), but since there are no unique links, this attempted paraphrase was classified as a *substantial revision*.

3.3.5.2 Stimulated recalls

The guideline for the retrospective interviews (see Appendix B) included passages that were used in this strand to gain further insights into the influencing variables that account for the quality of the written responses.

One of the introductory questions after the eye-tracking experiment asked for general comments on difficulties that participants encountered while working on the integrated writing task ("*Wie war das für Sie? Was war besonders schwierig? Was war*

leicht?“ / What was it like for you? What was particularly difficult? What was easy?). Idea units, i.e. segments that coherently commented on characteristics of the task like topic, complexity of the input material, or the expected genre, were assigned with the code “task variables”. This also applies for other passages in the interview where participants commented on the task.

In another question, participants were asked about their comprehension of the source material *(“Können Sie mir kurz zusammenfassen, worum es in dem Text und der Grafik ging?”/ “Could you briefly summarize what the source text and the graphical input were about?”*). Again, these passages in the stimulated recalls were divided into idea units that referred to single pieces of information in the input material. Each idea unit was then coded for the relevance and accuracy of information, applying the same codes that were used to code the written responses (see 3.3.5.1); codes were also assigned for the origin of information, i.e. whether the information was recalled from the source text or the graphical input. Results from this analysis, i.e. the comprehension of the source material, were then related to the analysis of the written performances.

3.3.5.3 Linking of process and product data

In a final step, the relationship between the quality of the written performances and viewing behavior of participants was explored.

Some researchers claim that through text analysis, and specifically by analyzing the text structure, cognitive processes could become evident in written products (Sanders & Schilperoord, 2006; Stezano Cotelo, 2003). Brinkschulte (2012) argues that writing processes are not per se apparent in the written texts:

> Leider findet ein Rückschluss von Textprodukten auf zugrundeliegende Schreibprozesse in Analysen zum akademischen Schreiben immer wieder statt. Was aus diesen Produkten geschlossen werden kann, sind einzig Schwächen in der sprachlichen und pragmatischen Ausgestaltung des Textes, jedoch nichts über dessen Entstehungsprozess. Deshalb müssten in Analysen von Schreibprozessen idealiter eine Korrelation von Prozess- und Produktdaten erfolgen. (Brinkschulte, 2012, p. 60) / [*Unfortunately, studies on academic writing make inferences from written products to the underlying cognitive processes. But what really can be concluded from the products are weaknesses regarding language and pragmatics, nothing about how these texts were produced. Therefore, the analysis of writing processes should include a correlation of process and product data.*]

The quality of the written responses was defined as the extent to which participants included relevant information and reproduced it correctly, as well as the extent to which they copied or transformed the language of the input material (as described in 3.3.5.1).

The assumption behind this linking of process and product data was, that a high dwell time in the input material together with frequent revisits to the source text and the graphical input would lead to (a) more relevant and correct information, and (b) to less copying from sources. Therefore, the results from content coding and the use of paraphrase types were correlated with the eye-tracking metrics of average dwell time (as a percentage of total task duration) and the total number of revisits in the AOIs source text and graphical input.

3.4 Findings

In this section, results from the analysis of the written performances will be reported to answer the overall research question of this product-oriented strand, namely: How do test takers process the information from the input material in the integrated

writing task of the digital TestDaF? And how do they transform the information from the two sources linguistically in their responses?

At the end of this section, findings from linking process data, i.e. eye-tracking metrics, to the quality of the written performances with respect to content and integration style will be presented.

3.4.1 Processing and transformation of input material

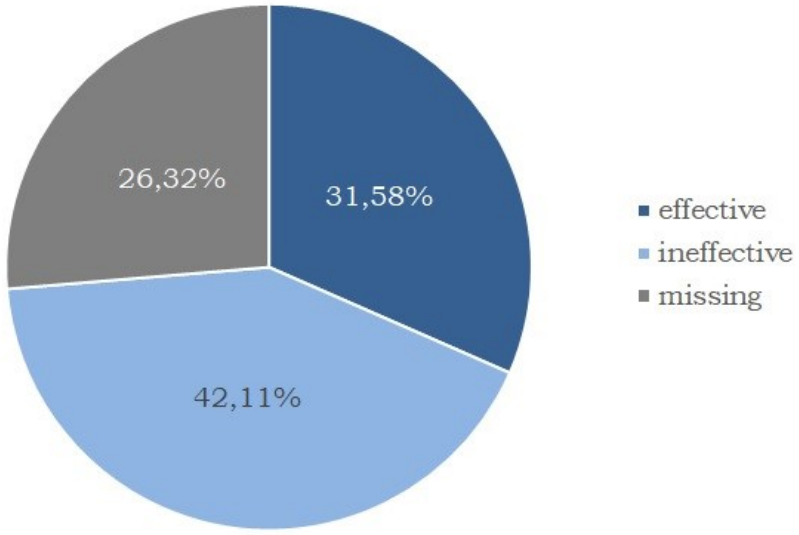
The main research question of this strand was to investigate how test takers process the information from the two sources with respect to content, and how they transform the input material linguistically.

In a first step, the average text length was calculated to see whether participants were meeting the requirements of the task. Participants wrote on average around 146 words (M=146.32; SD=32.24) during the 30 minutes to respond to the task, the shortest text was 106 words long, while the longest comprised 209 words and was beyond the expected word limit of 100 to 150 words.

Context statement

The written response to the integrated writing task of the digital TestDaF is intended to be a part of a written assignment at the university. The task does not explicitly require test takers to write an introduction, but nonetheless a short statement to introduce the subject would help the reader to contextualize the summarized information. The coding of this "context statement" (see Figure 3-3) showed that such an introductory statement was missing in five out of the 19 performances (around 26%). There were more ineffective introductions to the topic than effective context statements.

Figure 3-3 *Context statement*



Origin of information

Looking at the origin of information (see Table 3.4), the analysis showed that the performances on average contained more information from the source text than from the graphical input. There was a great variance between the individual responses. While some texts did not include any information from one of the two sources, others drew heavily on either the graphical input or the source text; in one case, the information included in the test takers response was solely taken from the reading input. Even though the summary is not meant to be balanced, i.e. to contain the same amount of information from the reading passage and the graphical input, the written responses should consider both sources and not only draw on one source. In addition, around 20% of the summaries also covered information that was not based on the sources. One performance consisted of very little information from the sources, but over 70% of the summary contained information that the participant added based on their background knowledge.

Table 3.4 *Origin of information*

	Graphical Input	**Source Text**	**unknown**
M (SD)	34.77 (19.29)	44.67 (29.36)	20.56 (21.31)
Mdn	42.86	44.44	11.11
Min.	.00	.00	.00
Max.	63.64	100.00	71.43

Note. N=19. Data are average percentages.

Source attribution

The coding of source attribution revealed that not all participants acknowledged the sources in their summaries (see Figure 3-4). In most of the cases where sources were acknowledged, the source attribution was ineffective (52%) rather than effective (24%).

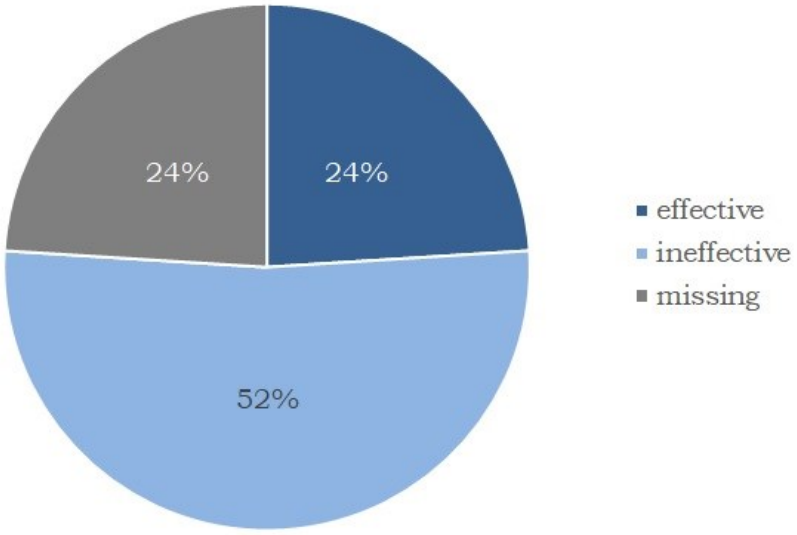
Examples of effective or ineffective acknowledgment of the sources from the collected performances can be found in Table 3.5.

Table 3.5 *Effective and ineffective examples of source attribution*

effective	Die Grafik mit dem Titel xxx informiert über… */The graph entitled xxx provides information about …*
	In dem Artikel xxx geht es vor allem darum zu zeigen, … / *The article xxx mainly wants to show …*
ineffective	Im Vergleich zu dem Text zeigt das Kreisdiagramm, dass … / *In comparison to the text, the pie chart shows …*
	Die Grafik zeigt einen Anteil von … / *The graph shows a proportion of …*

As these examples show, compared to the ineffective acknowledgment of the sources where participants simply referred to "the text" or "the graph", effective source attribution in addition provided the reader with more information about the sources by including the title of the source text and/or the graphical input.

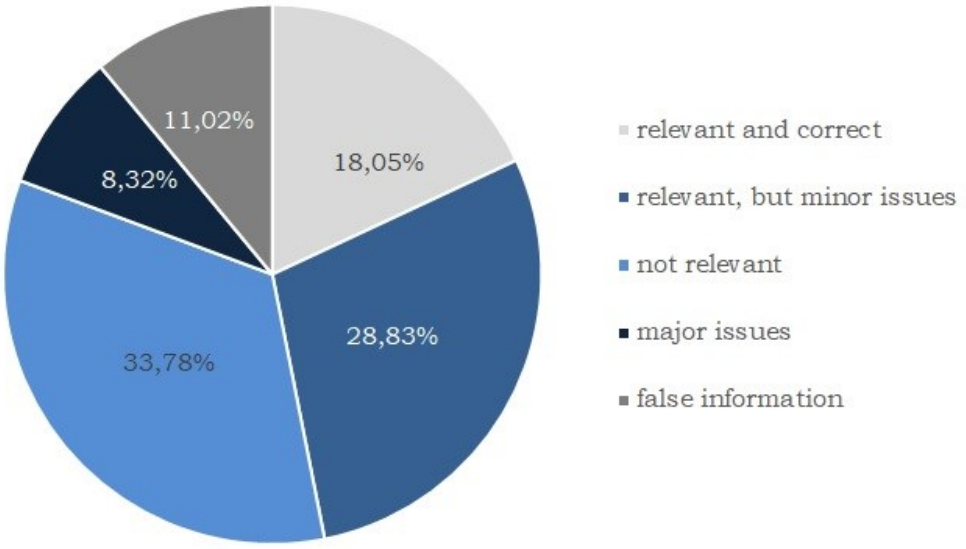
Figure 3-4 *Source attribution*



Relevance and accuracy of information

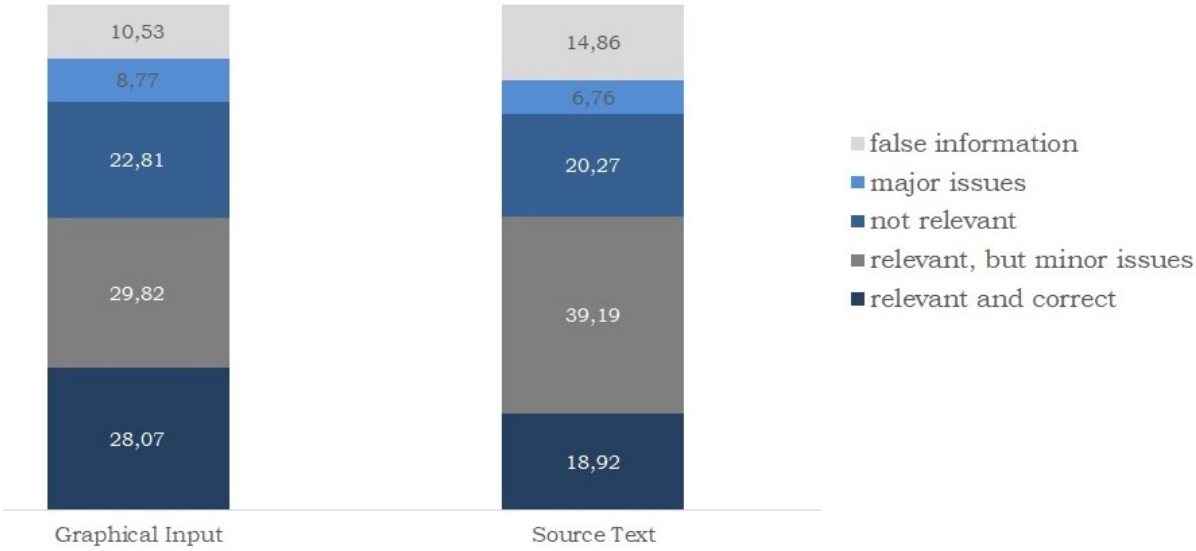
Taking a closer look at the relevance and accuracy of information in Figure 3-5, one can see that on average the summaries contained around 47% of relevant information, but only 18% were also correct, 28% were relevant, but had minor issues. More than half of the summaries consisted of information that was either not relevant, false, or was incomprehensible due to major language issues, whereby the proportion of not relevant information was the highest with around 34%.

Figure 3-5 *Relevance and accuracy of information (average percentage)*



To see whether the processing of information was affected by the type of source, relevance and accuracy was related to the origin of information. As shown in Figure 3-6, the proportion of relevant information from the graphical input and the source text was approximately the same with around 58%, though there were differences in the accuracy of information taken from the two sources. The percentage of relevant and correct information in the responses selected from the graphical input (around 28%) was higher compared to the relevant and correct information taken from the source text (approximately 19%). On the other hand, the percentage of relevant information with minor imprecisions was higher for the information originated from the source text (around 40%) than the information taken from the graphical input (about 30%).

Figure 3-6 *Proportion of relevant and accurate of information per source*



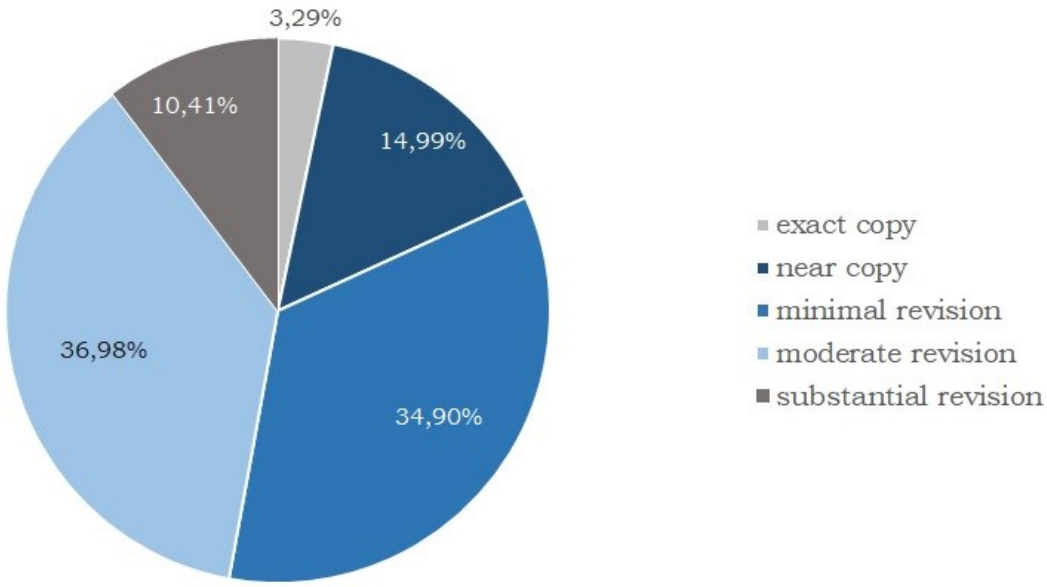
The analysis also showed that participants included about the same proportion of information from both sources which was not relevant with respect to the given question in the task instructions. The percentage of false information was higher for information taken from the source text (15%), only a small amount of reproduced information from both sources was not comprehensible due to major language issues.

Overall, one can see that participants drew on both sources in order to write their responses, but they also added information that was not based on the graphical input or the source text. They had difficulties selecting the relevant information and reproducing the ideas of the input material correctly. The relevance and accuracy of information included in the written performances also differed between the two sources, with more relevant information reproduced correctly from the graphical input than from the source text.

Integration style

Besides reproducing relevant information, the integrated writing task of the digital TestDaF also required participants to transform the language of the input material. The written responses were therefore also coded for different paraphrase types; the results are displayed in Figure 3-7.

Figure 3-7 *Paraphrase type (average percentages)*



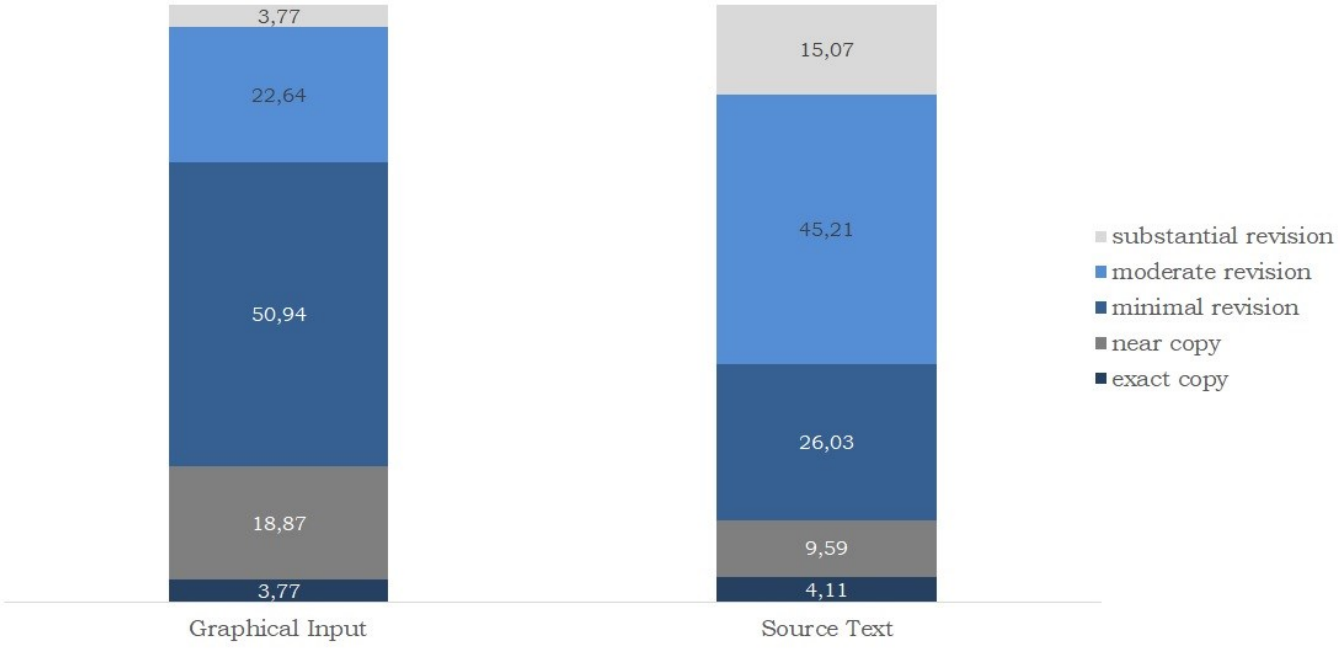
The analysis revealed that on average participants only copied a small amount of the input material verbatim (3%). A closer look at the exact copies in the performances showed that these were titles of the source text and the graphical input that participants either used as a heading for their own text, or incorporated them in their source attributions (see Table 3.5).

Besides these verbatim copies, still 50% of the wording of the summaries was very similar to language of the source material, with 15% near copies, i.e. “long copied strings taken from the original excerpt” (Keck, 2006, p. 268), and 34% of the texts only minimally

revised. The greatest proportion of paraphrase type used was moderate revision (37%), only 10% of the language material from the sources was substantially revised.

To see whether the transformation of language material differed between the two sources, the paraphrase type was related to the information reproduced from the graphical input and from the source text.

Figure 3-8 *Paraphrase type per source*



As Figure 3-8 shows, there were differences in relation to the transformation of language between the two types of sources. Participants only copied a small amount of language material verbatim from both sources (around 4%), but information taken from the graphical input was less transformed than information reproduced from the source text. The information participants included from the graphical input was phrased very similar to the original source, with over 50% of this information only being

minimally revised, and almost 19% near copies. In contrast, participants rephrased the information from the source text to a much greater extent. This becomes particularly apparent when looking at the proportion of moderate and substantial revisions: Moderate revisions were the most used paraphrase type for information reproduced from the source text, and the proportion of moderate revisions (45%) was the twice as high compared to information taken from the graphical input. The average percentage of substantial revisions was four times higher for information from the source text (15%) compared to information taken from the graphical input (almost 4%).

3.4.2 Effect of test-taker characteristics on the written performances

In this section, findings related to the effect of test-taker characteristics on the written performances are presented. To be more specific, the first part explores the role of test-taker characteristics like background knowledge and task representation from a qualitative perspective (RQ 1a), focusing on the comments participants made during the stimulated recalls that relate to these specific aspects. The second part investigates the impact of participants' overall language proficiency (as measured by the C-test) on the quality of the written responses (RQ 1b). Finally, the extent of which the comprehension of the input material affected the processing of the sources will be explored (RQ 1c).

Background knowledge and task representation

In the stimulated recalls, all participants commented on general difficulties they encountered when working on the integrated writing

task. Most of these comments were related to aspects that facilitated, or hindered the engagement with the task and the source material, specifically the background knowledge participants had on the given topic and the representation of the task requirements they were forming.

The difficulty of the topic was perceived differently, some participants said that the task was easy to work on, others reported on the lack of sufficient background knowledge that made it more challenging for them to work on the task, like e.g. participant 2-04:

> Ich hab keine Ahnung. Ich meine, ich hatte keine Vorkenntnisse. [*I have no idea. I mean, I didn't have any prior knowledge.*]

Later on in the interview, the participant addressed this issue again:

> Ja, ich habe, wie hab ich gesagt, ich hab kein Vorkenntnisse zu [XXX] überhaupt. Vielleicht wenn etwas ist über, weiß ich nicht, über Beziehungen oder so, das ist besser. [*Yes, as I said, I had no prior knowledge on [XXX]. Maybe, if it was about, I don't know, about relationships or so, that would be better.*]

As a major in psychology, participant 2-04 struggled with the topic of the task she worked on, and would have preferred a task from a topic domain that was more related to her own studies ("relationships").

Participant 2-03 also reported on challenges due to the lack of background knowledge:

> Oh, ich war überrascht, das ist sehr, es ist eine schwierige, schwieriges Thema für mich. Ich hatte keine Informationen darüber, und bis jetzt habe ich niemals diese Thema gelesen. Ja, und deshalb eine, ich bin schockiert ((lacht)) und dann muss ich zweimal oder dreimal eine, einen Satz lesen, gelesen. [*Oh, I was surprised. That was a difficult topic for me. I had no information about it, and until now I have never read about the topic. Yes, and therefore I was shocked ((laughs)), and then I had to read a sentence two or three times.*]

Apparently, the unfamiliar topic made it difficult for her to fully comprehend the source text, forcing her to read parts of it several times. Asked, if she had fully understood the reading passage by the end of task duration, she admitted, that this still was not the case.

Another aspect that participants mentioned in the interviews was the impact of task representation. As already reported in the previous chapter on cognitive processes, participants had difficulties with the task format since the newly developed integrated writing task was so different from their former writing experiences in preparatory language classes at university (see section 2.4.4). Apparently this did not only impact the cognitive processes, but also seemed to be a challenge for participants in shaping their understanding of the task requirements, and hence had an influence on the text they produced. For example, participant 1-06 described that he did not fully understand the task requirements and what he was supposed to do in order to summarize:

> Also, war für mich schwierig. Ich wusste nicht, was ich, äh, wie, also, zusammenfassen, aber Vorteile, Nachteile, Grund, Folgen, oder so was. Ich kannte nicht so genau, diese Sache, was wird von mir erwartet. Das war schwierig. [*So, for me it was difficult. I didn't know what I, how to summarize, but advantages, disadvantages, reason, consequences or something like that. I didn't know exactly, what was required from me. That was difficult.*]

The lack of sufficient background knowledge about the expected genre of summarization obviously led to a misinterpretation of the task requirements, and hence apparently influenced the written output:

> Und ich habe versucht eigentlich, den Text zu lesen, um ein bisschen davon Information bekommen. Aber die meisten Informationen habe ich von dem Grafik geschrieben. [*And I tried to read the text to get some information from there. But the most information I took from the graphic.*]

The participant acknowledged that he used the reading passage only as a secondary source for information, and that he based his summary mainly on the graphical input. This is also reflected in his written response to the integrated task: Only two out of the 11 sentences were based on information from the source text, seven referred to the graphical input, and another two sentences contained information not based on the source material.

Another example for such an unbalanced summary, i.e. a summary that is mainly based on one source, is the text that participant 1-09 produced: None of the sentences contained information from the graphical input, the whole text was based on information from the source text. During the interview, the participant explained that he did not include information from the graphical input in his summary because he thought it was not relevant for answering the task:

> Ich finde es, dass es nicht so viel gibt zu schreiben Ich finde es nicht so notwendig für dieses Thema. [*I think, that there is not much to write about. I think it is not necessary for the topic.*]

Even though the task instructions state that participants should summarize information in relation to the given question from both sources, the examples of participant 1-06 and participant 1-09 show that writers deviated from the intended task requirements, and created their own idea of what the written product should look like.

This conceptualization not only took place in the pre-writing phase, i.e. before participants started to write, but also throughout the entire writing process. Participants constantly evaluated their texts against the task instructions and their assumptions of the task requirements. Asked why he looked at his own writing and the instructions again even at the very end of the writing process, participant 2-05 answered:

> Dass ich unter Druck bin, ich weiß noch nicht, ob ich richtig beantwortet habe. [*That I'm under pressure. I don't know if I answered the question correctly.*]

Task representation was also influenced by participants' previous writing experience and acquired strategies. Participant 2-03, for example, wanted to write a conclusion at the end of her text, because that was something she had learnt:

> Aber ich habe gelernt, am Ende der Text muss ich eine Fazit von die Grafik... [*But I learnt that at the end of the text there has to be a conclusion.*]

At the end, she ended up deleting this part due to time constraints, but she conceptualized the written product in the integrated writing task to have a similar structure as the texts she had produced before.

In some cases, participants realized that their assumptions about the task requirements were wrong by closely reading the instructions again, as the example of participant 1-03 shows:

> Ja, ich, ich wollte schreiben „Ich bin der Meinung", aber meine Meinung geht hier nicht ((lacht)). [*Yes, I wanted to write "In my opinion", but my opinion won't work here ((laughs))*.]

As stated earlier, participants were enrolled in a preparatory language class for the paper-based TestDaF, and therefore made assumptions on the task requirements that were influenced by the format of the paper-based TestDaF, often relying on the structure and the formulaic sequences they had acquired to work on writing tasks that were more similar to opinion essays.

Since participants were not familiar with writing summaries, they also developed their own expectations about the underlying construct of the task. For example, participant 2-01 conceptualized the task as a vocabulary test:

> Und ich denke, diese Aufgabe versuchen unsere Wortschatz, Wortschatz zu prüfen. Ich möchte die verschiedene Wörter von andere Wörter zu ersetzen. Ich denke, das ist eine

> Zusammenfassung von dem Autor. Und ich muss eine andere Zusammenfassung machen. [*And I think that this task tries to assess vocabulary knowledge. I need to substitute the different words through other words. I think that this is a summary by the author. And I need to write a different summary.*]

The assumption of the integrated writing task being a vocabulary test might be based on the fact that participants have to use their own words to summarize the relevant information from the source material. Participant 2-01 therefore assumed that reformulating was one of the core requirements of the task. This conceptualization of the task affected her written response in that way, that she used a relatively high proportion of moderate (50%) and substantial (14.29%) revisions to reproduce the information from the input material.

To sum up, the qualitative interview data revealed that background knowledge on the topic had some effect on the way participants perceived the difficulty of the integrated writing task and that the conceptualization of the requirements affected the way they approached the task. To some extent, these influencing variables could be linked to the written products.

Overall language proficiency

In order to explore the effect of overall language proficiency on the processing and transformation of source material (RQ 1a), the sample was divided in low- and high-level learners according to their C-test results (see section 2.4.4). High-level learners were defined as participants who were placed at the CEFR-level B2 and above (N=8), participants with results below the threshold B2 were assigned to the group of low-level learners (N=11).

Non-parametric independent sample tests (Mann-Whitney U) were run to test for differences between the two groups for (a) the

origin of information, (b) the relevance and accuracy of information, as well as for (c) the paraphrase types participants used. Results are displayed in Table 3.6.

Table 3.6 *Comparison of low- and high-level learners*

	low (N=11)	**high** (N=8)	**Mann-Whitney U Test**		
	Mdn	***Mdn***	***U (Z)***	***p***	***r***
Origin of information					
Source Text	28.57	63.57	73.000 (2.403)	.016	.55
Graphical Input	44.44	31.67	23.500 (-1.705)	.091	
unknown	18.18	5.00	14.000 (-2.523)	.012	.58
Relevance and accuracy of information					
relevant and correct	20.00	17.79	48.500 (.378)	.717	
relevant, but minor issues	20.00	38.75	73.000 (2.403)	.016	.55
not relevant	42.86	26.79	13.500 (-2.535)	.009	.58
false	10.00	13.39	45.000 (.085)	1.000	
major issues	10.00	3.85	37.500 (-.568)	.600	
Paraphrase type					
exact copy	.00	.00	46.000 (.232)	.904	
near copy	16.67	.00	34.000 (-.875)	.442	
minimal revision	37.50	41.43	42.500 (-.125)	.904	
moderate revision	33.33	45.00	52.500 (.706)	.492	
substantial revision	.00	15.48	54.500 (.940)	.395	

With respect to content, the analysis of the written performances revealed that test takers processed the source material differently depending on their overall level of language proficiency. Weaker participants on average included mostly information from the graphical input (around 44%), while information included in the summaries of high-level learners was mainly taken from the source

text (63%). Also the proportion of information that was not based on the source material was much higher for the low-level participants (almost 18%) compared to those who had a higher level of language proficiency (around 5%). The differences varied significantly in relation to content originated from the source text and content not based on the input material, both with large effect sizes.

The relevance and accuracy of information also differed between the two groups. Low-level learners included more information in their summaries which was not relevant in relation to the given question, while the proportion of relevant information with minor imprecisions was the highest in the responses of high-level learners. Surprisingly, the percentage of relevant and correct information was a little higher in the responses of the low-level learners compared to the summaries of higher-level learners. The proportion of false information was about the same for both groups. The written performances of participants with a higher level of overall language proficiency also included some information that was not comprehensible, although the proportion was much smaller compared to the group of low-level learners. As can be seen in Table 3.6, only the differences for the relevant information with minor issues and for the not relevant information was statistically significant, but with large effect sizes.

Low- and high-level learners transformed the language of the input material differently. While the proportion of near copies as well as the percentage of minimal revisions was higher for participants with a lower level of language competence, the written responses of high-level learners demonstrated to a greater extent a linguistic transformation of the input material, as the proportion of moderate and substantially revised sentences shows. These differences in paraphrase types were not statistically different though.

Comprehension of sources

The focus of the third research question (1c) was to explore the effect of source comprehension on the processing of information in the written performances.

In a first step, excerpts from the stimulated recalls used for the eye-tracking experiment in Strand 1 (N=14) were analyzed. Therefore, participants' remarks related to source comprehension were coded for the relevance and accuracy of recalled information.

Table 3.7 *Source comprehension: Type of information in participants' stimulated recalls*

	relevant and correct	**relevant, but minor issues**	**not relevant**	**false information**
M (SD)	31.60 (25.51)	20.18 (15.56)	15.01 (18.53)	33.21 (35.33)
Mdn	30.95	21.11	2.78	29.17
Min.	.00	.00	.00	.00
Max.	80.00	50.00	50.00	100.00

Note. N=14. Data are average percentages.

As can be seen in Table 3.7, around one third of participants' recalled information from the sources was false, but almost the same amount of them were relevant and correct. Around 20% of information were reproduced with some imprecisions in the stimulated recalls. Participants also reported not relevant information, but in contrast to the written responses (see 3.4.1), there were no utterances in the interviews that were incomprehensible, i.e. were coded as having "major issues".

To see whether there were differences in comprehending the two sources, the relevance and accuracy of information were related to

the origin of information, i.e. whether participants mentioned information from the graphical input or from the source text.

Figure 3-9 *Source comprehension: Type of information per source*

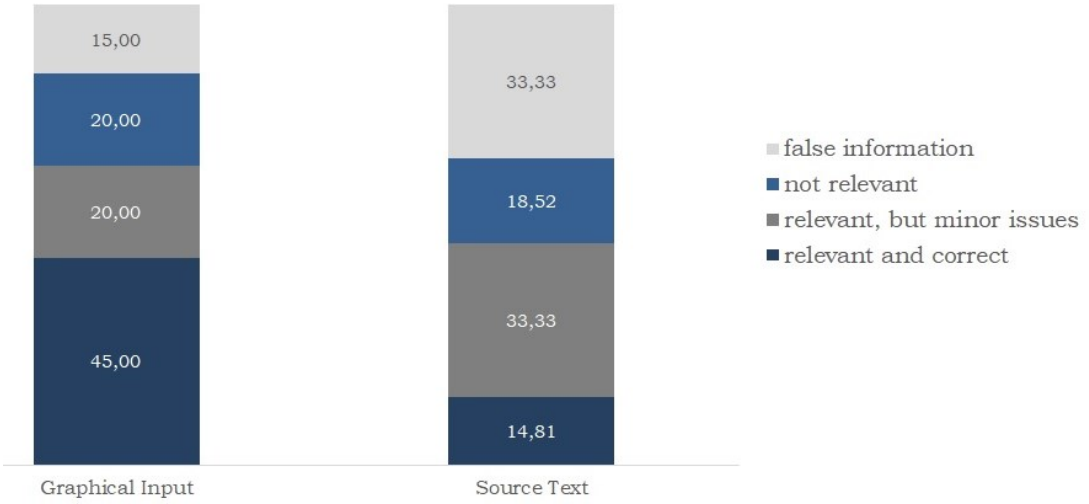


Figure 3-9 shows that participants referred to information from the graphical input more accurately compared to information from the source text. 65% of information recalled from the graphical input were relevant, 45% were reproduced correctly, while 20% had minor imprecisions. In contrast, one third of information related to the source text were false, and the proportion of relevant and correct information originated from the source text was relatively small with around 15%.

In a next step, the percentages of each type of information from passages in the interviews linked to source comprehension were correlated with the proportion of relevant and correct information in the written performances (see 3.4.1). The relationship between the comprehension of source material and the relevance and accuracy of information in the texts of participants was investigated using a

Spearman Rank Order Correlation (Spearman's rho).[34] Results of the correlation analysis are presented in Table 3.8.

Values for the correlation coefficient can range from -1 to +1, with values above 0 indicating a positive relationship between the two variables, meaning that as one variable increases, the other one increases as well. On the contrary, a negative relationship, indicated by values below 0, means that as one variable increases, the other variable decreases. A correlation coefficient of 0 indicates no relationship at all. Values between 0 and .3 indicate a weak, values between .3 and .7 a moderate, and values between .7 and 1.0 a strong relationship. The coefficient of determination, which is the value of r^2, is used to indicate the percentage of the variance in one variable that is explained by the other variable.

As can be seen in Table 3.8, there were large and statistically significant correlations between the percentage of relevant information participants reproduced in the interviews correctly and the proportion of relevant information they included in their written texts. This finding suggests that there is a relation between source comprehension and the successful processing of information in the TestDaF integrated writing task, and is supported by the large negative correlation between relevant and correct information recalled by participants in the interviews and the proportion of not relevant and false information in the summaries. This means that participants who correctly referred to relevant information from the sources in the stimulated recalls, were also able to reproduce this information in their texts, and did not include information which was not relevant or false.

[34] Preliminary analysis revealed that not all data was normally distributed (Shapiro Wilk, $p < .05$).

The correlation analysis also revealed a strong correlation between information that was not correctly reproduced in the retrospective interviews and the proportion of false information participants included in their texts, again implying an effect of comprehending the sources on the processing of input material. Participants who did not recall the information from the sources correctly were likely to integrate wrong and not relevant information in their response to the task.

Table 3.8 *Spearman rank order correlation between source comprehension and relevance and accuracy of information in the written performances*

		Written performances				
	type of information	relevant and correct	relevant, but minor issues	not relevant	false	major issues
Source comprehension	relevant and correct	.636*	.563*	-.844**	-.555*	-.128
	relevant, but minor issues	.372	.071	-.080	-.197	.-062
	not relevant	-.223	.263	.090	-.165	-.018
	false	-.494	-.592*	.641*	.549*	.149

Note. N=14. * p < .05. ** p < .01

3.4.3 Generalizability of results

Another research question of this strand wanted to look into possible task effects, i.e. whether the processing of information and linguistic transformation of the input material can be generalized across different versions of the task, i.e. Set 1 and Set 2 (RQ 1d). To test for significant differences between the two sets, non-parametric independent samples Mann-Whitney U tests were used to look into (a) the source of information, (b) how relevant the information was in relation to the given question and how accurate participants

reproduced this information, as well as (c) to what extent they rephrased the language of the input material. Results of this analyses are displayed in Table 3.9.

As can be seen, participants processed the information from the two sources differently depending on the task version they were working on. Summaries from participants in Set 2 contained more information from the source text, while the proportion of information originated from the graphical input and of additional information not based on the sources was higher in Set 1.

Table 3.9 *Comparison of Set 1 and Set 2*

	Set 1 (N=11)	**Set 2** (N=8)	**Mann-Whitney U Test**	
	Mdn	***Mdn***	***U (Z)***	***p***
Origin of information				
Source Text	33.33	50.00	55.500 (.953)	.351
Graphical Input	44.44	37.50	30.000 (-1.164)	.272
unknown	17.43	11.11	31.500 (-1.051)	.310
Relevance and accuracy of information				
relevant and correct	18.34	22.22	49.000 (.420)	.717
relevant, but minor issues	28.64	28.57	39.000 (-.414)	.717
not relevant	36.67	33.33	43.000 (-.083)	.968
false	10.00	14.29	48.000 (.339)	.778
major issues	5.00	7.69	50.000 (.525)	.657
Paraphrase type				
exact copy	.00	.00	49.000 (.579)	.717
near copy	.00	14.29	53.000 (.787)	.492
minimal revision	38.09	40.00	36.500 (-.625)	.545
moderate revision	29.17	50.00	60.000 (1.328)	.206
substantial revision	18.33	.00	25.000 (-1.701)	.129

The differences related to the relevance and accuracy of information existed, but were small between the two groups; participants in both sets mostly included not relevant information in their summaries, followed by information that was relevant, but had minor issues, i.e. were not precisely reproduced. Only a small percentage of information in the written performances of both sets was not comprehensible due to language problems.

Looking at how participants transformed the language of the input material, the analysis revealed variances across paraphrase type and task versions, but with no clear pattern emerging. Summaries from participants assigned to Set 1 contained less near copies than the written performances of participants from Set 2. On the other hand, the proportion of substantial revisions was higher for this group, while the participants in Set 2 used moderate revisions.

Though there were differences between the two different task versions with regard to the processing and transformation of the input material, none of them was statistically significant.

3.4.4 Linking of process and product data

The aim of the final research question (RQ 2) was to see whether the inclusion of relevant information and the transformation of the input material in the written performances was linked to the viewing behavior of the participants. To explore a possible relation of process and product data, correlation analysis (Spearman's Rho) was used.

As the results presented in Table 3.10. show, there was a strong and statistically significant correlation between the total number of revisits in the AOI source text and the relevance and accuracy of information ($r = .610$, $p < .01$), as well as the percentage of

substantial revisions (r = .610, p < .01). This means, that participants who went back to the reading passage more frequently during task completion included more relevant and correct information in their responses. This viewing behavior also led to a higher percentage of substantial revisions, i.e. participants reproduced the information with their own words instead of just simply copying long strings of words from the input material.

Table 3.10 *Spearman rank order correlation between quality of the written performances and viewing behavior*

		average dwell time		**total number of revisits**	
		Source text	Graphical input	Source Text	Graphical Input
Content Coding	relevant and correct	.184	-.010	.610**	.108
	relevant, but minor issues	.256	-.343	.090	-.254
	not relevant	-.412	.027	-.490*	-.038
	false information	.060	.173	-.097	-.023
	major issues	-.029	.254	-.152	.199
Paraphrase Type	exact copy	.076	.096	.201	-.287
	near copy	.101	-.252	.061	-.501*
	minimal revision	-.233	.560*	-.275	.513*
	moderate revision	-.061	-.601*	-.048	-.336
	substantial revision	.437	-.010	.610**	.108

Note. N=17. * p < .05. ** p < .01

The shared variance, which is the coefficient of determination, indicate that around 37% ($.610^2 * 100$) of the information coded as relevant and correct can be explained by the total number of revisits.

The same applies to the shared variance of substantial revisions and the number of revisits in the AOI source text.

The average dwell time in the AOI source text apparently did not have a statistically significant effect on integration style. Only the time participants spent in the AOI graphical input ($r = .560$, $p < .05$) and the frequency in which they went back to this specific AOI ($r = .513$, $p < .05$) were related to the paraphrase type of minimal revision. This paraphrase type was also the one most frequently used for reproducing information from the graphical input (see Figure 3-8), which could also explain to some extent the relation between the viewing behavior and the linguistic transformation of the information from this source type.

Overall, the hypothesis that a high dwell time in the input material together with frequent revisits to the sources would impact the quality of the written responses could only be partly confirmed. Especially the number of revisits to the source text seemed to be related to the percentage of relevant and correct information in the written responses, and a more comprehensive transformation of the language of the input material.

3.5 Discussion

The overall research aim of this strand was to investigate how test takers process the information from the input material in the integrated writing task of the digital TestDaF, and in what way they transformed the language material from the two sources linguistically.

As expected, findings showed that test takers' responses relied to a great extent on the source material. Contrary to expectations

regarding summary writing, they also included information that was based on their own background knowledge. It seems as if this was mainly related to a misinterpretation of the task demands, as expressed by some participants in the stimulated recalls. One of the issues emerging from this is the need for test preparation material that makes the task demands transparent to potential test takers. As mentioned earlier (see section 1.5.3), data was collected at an early piloting stage, and participants had almost no opportunity to familiarize themselves with the requirements of the new test format. As a response to test taker feedback collected via questionnaire during piloting, short tutorials for every single one of the 23 test tasks of the digital TestDaF were produced, informing prospective test takers about the task demands. These short videos are now publicly available on the TestDaF website together with a model test and other information material.[35]

Results from this study are consistent with previous research that has identified language proficiency as a determining factor in the quality of integrated writing performances (e.g. Chin, 2009; Keck, 2006; Plakans & Gebril, 2013; Shi, 2004). The comparison of low- and high-level learners in this study revealed differences with regard to a) the relevance and accuracy of information included in the summaries, b) the type of paraphrase participants used to reproduce the information, and c) the source type the information was taken from. Low-level learners included less relevant, but more irrelevant information in their summaries. They also used more near copies and minimal revisions, and based their summaries more on information from the graphical input. The unbalanced summaries of low-level learners were probably related to problems with the comprehension of sources, as reported in the chapter on cognitive

[35] https://www.testdaf.de/de/teilnehmende/der-digitale-testdaf/vorbereitung-auf-den-digitalen-testdaf/

processes in the current study (see section 2.4.4) and as in Plakans and Gebril (2017). On the contrary, performances of high-level learners were characterized by more relevant information, the use of more moderate and substantial revisions, as well as a higher proportion of information taken from the source text.

The analysis of the written performances also demonstrated that differences regarding the relevance and accuracy of information as well as paraphrase type were not only related to the overall language proficiency of participants. They were also related to source type, i.e. there were differences between information reproduced from the source text and the graphical input. This issue has not been addressed in integrated writing research yet. Results showed that the proportion of relevant information was about the same for both sources, but the information from the graphical input was reproduced more accurately compared to information from the source text. One explanation could be that the possibility to reproduce relevant information incorrect or imprecise appears to be lower for the graphical input compared to the source text since the amount of information and the cognitive load for processing differ between the two source types. Another explanation might be related to the use of paraphrase types that was different for the two sources. The analysis revealed that language material taken from the graphical input was only minimally revised, whereas the reproduction of information from the source text included more moderate and substantial revisions. This seems logic, given the fact that verbalizing information from graphics only requires the processing of relevant data points and short strings of words as in the title or the graphics legend. Hence, it could be argued that there is hardly no other way to rephrase information from a graphical input without using the language material that is already included in the graphic itself.

In general, it might be worth reconsidering paraphrasing as an indicator for L2 integrated writing quality. Research has shown that L2 writers tend to do more copying instead of paraphrasing compared to L1 writers (Grabe & Zhang, 2013; Keck, 2006; Shi, 2004). And especially summary writing as a specific kind of integrated writing is characterized by more textual borrowing and less lexical diversity (e.g. Shi, 2004; Yu, 2013b).

The quality of the written performances was only weakly linked to the cognitive processes reported in Strand 1. Only the number of revisits to the source text seemed to be related to the percentage of relevant and correct information in the written responses, and a more comprehensive transformation of the language of the input material. Claims about processes on the basis of products should therefore be cautioned

3.5.1 Limitations

The written responses analyzed in this strand were from the same sample which participated in the eye-tracking experiment of Strand 1 to allow for linking process and product data. Therefore, the same sample-related limitations are valid: (1) small sample size, (2) relatively low language proficiency, and (3) possible washback effects from preparing for the paper-based version of the TestDaF (see section 2.5.1 for more details).

One possibility to overcome these limitations in further research on the quality of integrated writing performances in the digital TestDaF, would be to increase the sample size. This would probably limit the chance to relate product and process data since a larger sample size for investigating the cognitive processes within a mixed-methods design is unlikely (see section 2.5.1). But the analysis of a large number of texts would provide further insights into the quality

of test takers' written responses, particularly with regard to text quality across different levels of writing proficiency. In addition, it would be beneficial for the analysis if the written responses came from live examinations, i.e. from test takers who intentionally prepared for the digital TestDaF to determine the effect of task familiarity on the written responses which was observed in the current study.

In relation to the research methodology, the following limitations should be considered for the interpretation of findings: Unlike in Keck (2006), *general* and *unique links* were determined manually in the current study, so that the consistency of applying the same criteria across all texts analyzed might be questioned. This also has implications for the coding of paraphrase type, since *unique links* were used as a basis for establishing the taxonomy of paraphrase types. Keck's classification of paraphrase types was based on written input material. The current study revealed differences in integration style depending on the type of source, i.e. graphical input and reading passage, and showed that information reproduced from the graphical input was to a great extent only minimally revised. The validity of the paraphrase type taxonomy with regard to input other than written sources could therefore be questioned.

In addition, the coding of test takers' responses was focused on content and integration style, other factors that are usually associated with quality of writing, like for example syntactic complexity, range or correctness were not taken into account.

To look into source text comprehension, parts of the stimulated recall data collected for Strand 1 was used. Participants were asked if they could summarize what the text and the graphical input were about ("*Können Sie mir kurz zusammenfassen, worum es in dem Text und der Grafik ging?*"; see also Appendix B). A note of caution is due here since this question is very generic, not specifically asking for

the relevance of information, whereas the excerpts of the stimulated recalls were coded with the same coding scheme that was used for coding the written performances (Figure 3-1).

Even though in this study no quantitative data was included to measure the effect of background knowledge on the written responses, the qualitative interview data still provided some insights into how participants perceived the difficulty of the topic they were working on. The unfamiliarity with the topic might have had an impact on the comprehension of the sources, as well as the ability to reproduce the information in their own words.

4 SCORING OF INTEGRATED WRITING PERFORMANCES

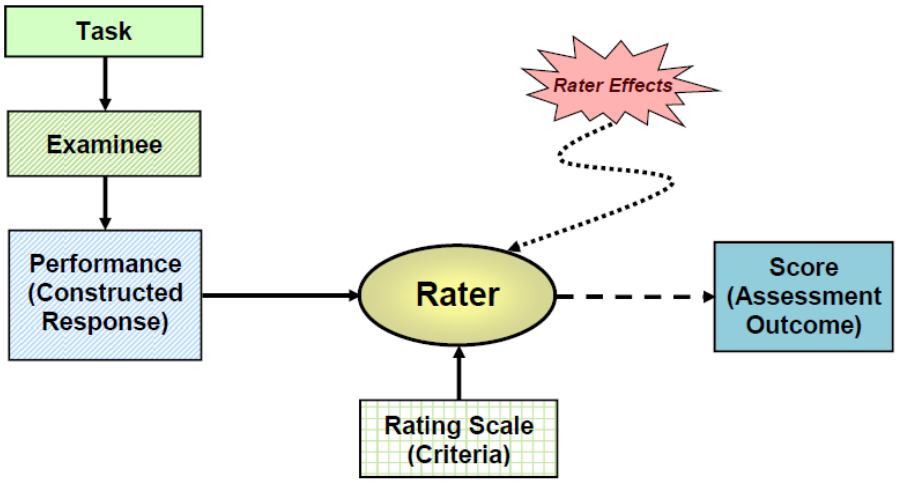
The previous chapter approached the construct underlying the TestDaF integrated writing by taking a closer look at the quality of test takers' responses in relation to source use and integration style. The following chapter covers the final strand of the mixed-method study. Statistical analyses are used to a) examine the reliability of rating integrated writing performances, and b) to explore reading-writing relations to provide evidence for the underlying assumption that the quality of the written performances is reflected in the score (see the overall research aim in the Introduction).

The first section of the chapter gives a brief introduction into issues in rater-mediated writing assessment. Section 4.2 reviews existing research that has looked into the relation of reading and writing from the perspective of scoring integrated writing performances. The methodology used in this strand is outlined in section 4.3, before findings are presented in the subsequent section 4.4. The chapter closes with a discussion of results.

4.1 Reliability in rater-mediated writing assessment

In rater-mediated or performance-based assessment, test scores are based on a performance of an examinee (e.g. a written response to a constructed-response item), and the associated rating of this performance by a rater (see Figure 4-1). Research has shown that scores in rater-mediated assessment are confounded by a number of factors such as tasks, rating criteria, and rater variability (e.g. Eckes, 2005; McNamara, 1996). As a result, score reliability in writing assessments is lower compared to other assessment types (Gebril, 2010).

Figure 4-1 *Basic structure of rater-mediated assessment*[36]



Rater effects pose a threat to the validity of the assessment procedure since they are irrelevant to the construct being measured. (Bachman, 2004; Weir, 2005). Usually test providers approach this issue through rater training and monitoring, but studies have shown that these measures do not necessarily improve rater reliability (e.g. Elder, Knoch, Barkhuizen & Randow, 2005; Lumley & McNamara, 1995).

One way to account for variability in terms of rater effects like severity/leniency or central tendency in the assessment procedure is the application of a specific measurement model: *many-facet Rasch measurement* (MFRM; Linacre, 1989). In contrast to *generalizability theory* (G-theory; Brennan, 2001), another measurement model that "constitutes a theoretical framework and set of procedures for specifying and estimating the relative effects of different factors on test scores" (Bachman, 1997, p. 255), MFRM analyzes the different factors, or facets simultaneously. These facets,

[36] Courtesy of Dr. Thomas Eckes, TestDaF-Institut, Ruhr-University, Bochum, Germany. Figure is also included in Eckes (2019).

e.g. raters, writing tasks, or rating criteria are then mapped onto a single linear scale, the logit scale. The joint calibration of facets allows for addressing issues in relation to the rater facet, e.g. whether raters consistently assign scores too low or to high (rater severity/leniency), or how consistent they are (Barkaoui, 2014b).[37]

With the increasing use of integrated writing tasks, there is also a growing need to look into sources of measurement errors related to this specific task type since integrated writing assessment introduces new challenges for raters, as Wang, Engelhard, Raczynski, Song, and Wolfe (2017) showed in their study on rater accuracy and perception. Not only do raters have to assess the writing ability of test takers, they also have to rate the ability to read and comprehend the sources by evaluating the incorporation of relevant information in the written performances. Wang et al. used difficult-to-score essays to examine scoring decisions of raters. The qualitative analysis of the applied rater perception survey revealed that raters had different perceptions of three major essays features, i.e. the focus of the essay, textual borrowing, and idea development.

Challenges in rating integrated writing performances were also reported by Gebril and Plakans (2014). Their inductive analysis of the rating process of two raters revealed that it was difficult for raters to distinguish between the language from the sources and language produced by the writer. Problematic were also essays that either contained a high number of quotations instead of paraphrases, or essays that copied from the input material without crediting the sources.

While these qualitative findings show that issues related to source use and integration style represent a challenge for raters, the

[37] For a comprehensive overview to MFRM see Eckes (2015a); (2019).

reliability of integrated writing scores has not been sufficiently investigated (Shin & Ewert, 2015).

4.2 Reading-writing relations in integrated writing scores

As shown previously in this study, integrated writing requires writing as well as reading ability. One question though remains: What accounts for the integrated writing score? To answer this question, research has looked into the impact that reading and writing have on integrated scores, with mixed findings.

Watanabe (2001) correlated test takers' integrated writing performances with a reading test and a writing-only task. The results confirmed the central role of writing in reading-to-write tasks, and showed only weak correlations between the reading test and the integrated task.

Similar results in relation to the weight of reading in integrated writing were also reported by Asención Delaney (2008) who explored the construct underlying two different types of integrated writing, a summary task and a response essay. A correlation analysis of these two tasks with both a writing and a reading test revealed only a weak correlation with the reading test. But unlike Watanabe (2001), Asención Delaney did not find a strong correlation between the integrated tasks and the writing measure. In addition, her study also showed only a weak correlation between the two integrated task types, implying "that the performance on the summary and the response essay tasks could be considered two different dimensions of the reading-to-write ability" (Asención Delaney, 2008, p. 144).

These findings are different from Shin and Ewert (2015). Their study on the development of an analytic rubric for rating integrated writing performances of college ESL students found a moderate correlation between a reading measure and a reading-into-writing

task. Since the correlation with a writing measure was similar, Shin & Ewert suggest that reading and writing are central components of integrated writing ability, and that integrated writing tasks "may tap into both reading and writing ability" (ibid., p. 15).

Sawaki, Quinlan, and Lee (2013) used a large-scale factor analysis to analyzed responses to the TOEFL iBT integrated writing task. The aim of the analysis was to investigate the construct underlying the integrated writing task (that integrates reading, listening, and writing) and how this is related to reading and listening comprehension. Results "revealed the identification of three correlated and yet distinct constructs" (ibid., p. 92), i.e. reading, listening and writing.

4.3 Methodology

4.3.1 Research aims and questions

The aim of this strand is twofold. The first main research question focuses on the rating of the integrated writing performances:

RQ 1: Are the integrated writing performances rated reliably? Are the results generalizable across task versions?

The following sub-questions should provide further insights into the rating of the TestDaF integrated writing task:

RQ 1a: Can the rating scale reliably distinguish integrated writing performances at different levels?

RQ 1b: Can raters apply the rating criteria reliably and consistently?

RQ 1c: Is the quality of the written performances reflected in the score?

The second research question explores possible relations between integrated writing performances and reading or writing ability of test takers.

RQ 2: Does the integrated writing task of the digital TestDaF measure writing or reading competence? To be more specific: Are the results on the integrated writing task related to independent writing scores, or to scores yielded in a reading comprehension test?

4.3.2 Participants

Test takers

Data from piloting the new test format in test centers across the world was used in this strand. The piloting included two different task versions of the whole test, each version (Set 1 or Set 2) was randomly assigned to one of participating test centers.

Overall, 445 participants took part in the piloting, but only participants where results from the C-test, scores from the reading component, ratings from the writing component and demographic data was available were included in analysis.

As can be seen in Table 4.1, participants in both sets were on average around 27 years old, with ages ranging from 17 to 61.[38] The majority of them were women, with the proportion of female participants being higher in Set 1 compared to Set 2. The regional distribution differed between the two groups as the list of the main countries of origin shows, but the samples are still representative for the expected TestDaF population.[39] The results of the accompanying

[38] The relatively high age of some participants could be related to the fact that some teachers enrolled in the piloting to become familiar with the new test format.
[39] For a list of learners who have taken the TestDaF in the last years, refer to the statistical overview of "Compact Data" which can be found on the TestDaF website (www.testdaf.de; available in German and English).

C-test (see 1.5.2) revealed that the overall language competence of the participants was relatively low. A majority of participants yielded results below the required threshold level of B2 (Set 1: 56.5%, Set 2: 70.4%).

Table 4.1 *Characteristics of participants across task versions*

	Set 1	**Set 2**
Number of participants		
Reading	215	180
Writing	223	159
Age		
M (SD)	26.70 (7.40)	27.20 (7.57)
Min.	17	19
Max.	59	61
Gender		
female	62.8%	52.8%
male	37.2%	47.2%
Main countries of origin	Russia: 15.2% Syria:13.5% Hungary: 12.1% Cameroon: 9.4% Taiwan: 7.6% Turkey: 7.6% China: 7.2%	Iran: 21.4% China: 19.5% Russia: 13.8% Tunisia: 8.8% South Korea: 6.9% Brazil: 6.3% Turkey: 5.7%
C-Test results		
under A2	4.0%	6.3%
A2	15.7%	17.6%
B1	36.8%	46.5%
B2	33.7%	20.7%
C1 and above	9.9%	8.8%

Raters

The written responses were rated by 28 experienced raters, two male and 26 female raters. 25 of them were external raters with many years of experience in rating test takers' responses from the paper-based TestDaF.[40] The other three were internal language testing specialists who were involved in the development of the

[40] The prerequisites for becoming a certified rater for TestDaF performances and the qualification process are described on the TestDaF website (www.testdaf.de; in German only).

digital TestDaF and the according rating scales for writing and speaking. All external raters received an intensive two-day training in applying the newly developed rating scales to the written and spoken performances.

4.3.3 Instruments and procedure

Data was collected during piloting new test material for the digital TestDaF in June 2018. Piloting took place in 39 licensed test centers across the world. Participants had access to a model test beforehand to become familiar with the newly developed task types.

All participants had to sit a C-test (see 1.5.2) before working on the reading and writing component of the digital TestDaF.

Reading comprehension test

The reading component of the digital TestDaF assesses to what extent international study applicants and speakers of German as a foreign language can comprehend written texts that are relevant in the academic context. The construct of the reading component initiates underlying cognitive processes that are involved in reading for different purposes at the university, e.g. finding information, comprehending texts, reading to learn, reading to integrate information (Enright et al., 2000). Based on this, the tasks amongst others assess if a test taker is able to understand the structure and the main points of a text, to recognize causal relations like reasons and consequences, or to compare and contrast information from different sources. The tasks require a broad linguistic range, including pragmatic knowledge.

The reading component consists of seven tasks with 34 closed items in different item formats (e.g. multiple-choice, ordering,

highlighting false information on a sentence level). The time for each tasks varies between 4 and 15 minutes, overall the reading component lasts approximately 60 minutes.

Independent writing task

The writing component of the digital TestDaF consists of two tasks: one integrated writing task (see 1.2), and one independent writing task.

The independent writing task is a typical *timed impromptu essay* (Eckes, Müller-Karabil, & Zimmermann, 2016; Weigle, 2002) in which candidates have to take a position to a given topic, discuss positive and negative aspects or advantages and disadvantages respectively by presenting relevant arguments that are supported by justifications and examples. The task assesses test takers ability to write a coherent, discursive text. The task requires an extensive planning phase to generate language and content from scratch, building on test taker's own background knowledge, as well as linguistic range and grammatical competence. The task instructions include an open question to a given topic, optionally one or two very short statements (one to two sentences long) are added. Test takers have 30 minutes to complete the task by writing a minimum of 200 words.

Rating scales

After the piloting phase, the written performances on the independent and the integrated writing task were scored on 6-point holistic scales (0-5 points), one for each task type. The scales take into account the specific nature of each task, e.g. the rating scale for the integrated task addresses the extent to which the relevant

information from both sources have been selected and reproduced correctly, and also evaluates the linguistic transformation of the input material (Chan et al., 2015; Knoch & Sitajalabhorn, 2013). Rating was done onscreen, and during rating, raters had access to the performances, the rating scale, and a clear description of what could be expected from the performances in the online rating tool. The latter comprised a list of relevant information from the sources that should be included in the summaries, and with details about the anticipated linguistic range of the written texts.[41]

4.3.4 Data analysis

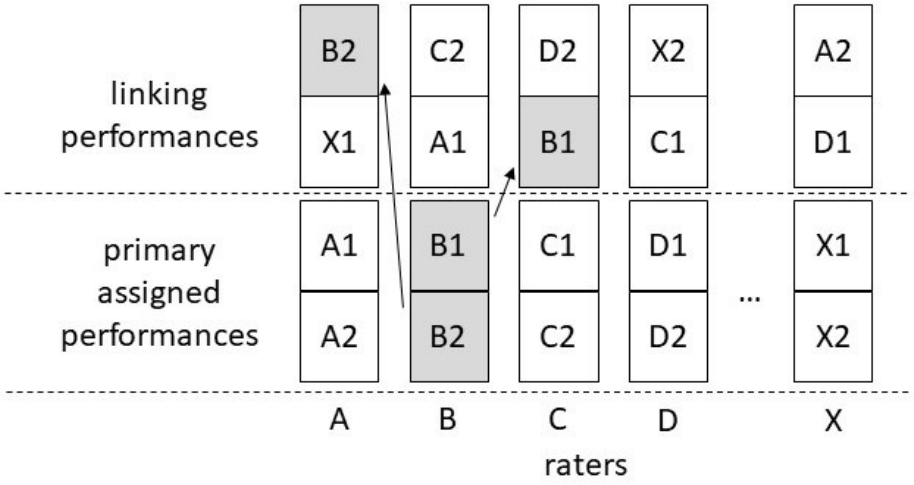
Data from piloting new test material is routinely analyzed in the department of Psychometrics at the TestDaF-Institut applying MFRM. The data used in the current study was analyzed with FACETS, Version No. 3.80.1.

Each written performance of both the integrated and the independent writing task was rated independently by two raters. The high number of collected data points should yield higher measurement precision of examinees proficiency estimates. To guarantee the connectedness of the data set, an incomplete linked design was chosen (see Figure 4-2). Placed in a random order, raters were assigned ten primary performances. In addition, each rater also had to rate ten linking performances from other raters. For example, in Figure 4-2, rater B was assigned the primary performances B1 and B2. B1 was also assigned to rater C which he/she had to rate in addition to his/her primary performances (C1 and C2), B2 was a linking performance for rater A. Though this linking design, all

[41] For rating written performances from actual live tests, raters additionally receive calibrated benchmark performances taken from the piloting, illustrating the different levels of the rating scale for each specific task version.

raters had to rate 20 performances in total, but all performances were rated by two raters. The results from MFRM were used to look into the reliability of ratings within and across test versions (RQ 1a and RQ 1b).

Figure 4-2 *Rating design*



The 19 written responses which were analyzed in Strand 2 regarding content and integration style were used to investigate whether the quality of the integrated writing performances was reflected in the score (RQ 1c). For this purpose, the sample was divided in low- and high-scoring participants according to their fair average they received in the integrated writing task (see Appendix G). High-scorers were defined as participants who yielded a fair average of ≥ 2.50 logits, participants with a fair average below that threshold were considered as low-scorers. According to this distinction, only five out of the 19 participants were high-scoring participants, the remaining 14 were considered low-scoring participants. Non-parametric independent sample tests (Mann-Whitney U) were run to test for differences between the two groups regarding (a) the origin of information, (b) the relevance and accuracy of information, as well as for (c) the paraphrase types used.

And finally, a correlation analysis was used to explore how integrated writing performances was related to the reading comprehension test and the independent writing.

4.4 Findings

4.4.1 Reliability of ratings

One of the overall aims of this strand was to look into the reliability of the rating of integrated writing performances across task versions (RQ 1).

Useful insights into the results from piloting and rating the new test material across task versions were provided by W*right maps*. Separation statistics and scale functioning analysis were used to explore whether the task was able to discriminate between distinct levels of examinees' writing proficiency, and to investigate whether the rating scale functioned (RQ 1a). The reliability and consistency of raters in applying the rating scales (RQ 1b) were investigated with the help of rater measurement reports.

4.4.1.1 Wright Map

A *Wright Map* (see Figure 4-3 as an example) is part of the analysis output of the FACETS program which allows for a direct comparison of examinees, raters and tasks (Eckes, 2015a). All three facets have been calibrated onto one single measurement scale.

- The first column shows the common Rasch scale (*Measure*) with logits being the measurement units.
- The column *Examinees* displays examinees proficiency estimates. Examinees are represented by stars and dots: a star stands for two participants (indicated in the bottom line of the column), a dot means one participant. *Examinees* is a positively measured

facet (indicated by the "+" in the column title), meaning that a high proficiency is represented by a high score; conversely, low scores represent a low proficiency. This means that more proficient examinees are placed at the top, less proficient examinees at the bottom of the column.

- The third column compares raters in terms of their severity in rating performances. Each rater is represented by an individual rater ID. *Rater* is a negatively oriented facet (indicated by the "-" in the column title). This means that higher severity measures result in lower scores assigned to examinees, while low severity measures result in higher scores awarded to examinees. In other words: The higher a rater is located in the column, the more severe he/she is.
- The column *Tasks* shows the difficulty measures for the tasks under consideration. This measure is also negatively oriented. A high measure indicates that the task is difficult, meaning that even proficient examinees are less likely to receive a high score in this task. The lower a task is located in the column, the easier it is for examinees to yield higher scores.
- The last column in the Wright Map (*Scale*) maps the 6-point rating scale to the logit scale. There are horizontal dashed lines in this column, representing the *category thresholds*, i.e. the boundaries between the scores on the rating scale.

Set 1

As can be seen in the Wright Map for Set 1 (Figure 4-3), the proficiency estimates of the examinees varied to a great extent. The logit spread for the proficiency measure was around 22 logits, with some high-proficient examinees (above 10.00 logits), and some very low-proficient examinees (almost -12.00 logits). A high proportion of

examinees were centered around the middle, with a majority of them above the mean element measure of zero.

Looking at the column *Raters*, one can see that there was quite a variation amongst raters regarding their level of severity. Rater 901 is the most severe rater with a severity measure of about 3 logits, while raters 179 and 222 were the most lenient raters with a severity measure of around two-and-a-half logits. With rater severity estimates ≥ 1.0 logits or ≤ -1.0 logits being considered as "severe" or "lenient" (Eckes, 2019), one can see from the Wright Map that there was also a high number of raters who did not tend towards rating more harshly or more leniently.

If the distribution of examinee proficiency corresponds with rater severity and task difficulty on the logit scale, it is considered that the task is not too difficult or too easy for the specific population of test takers under consideration. Looking at the Wright Map for Set 1, one can see that this is the case for the integrated writing task in Set 1. More information about task difficulty can be gained by including the two different task types of the writing component in the analysis (see Appendix H). As can be seen from the Wright Map for both sets, the integrated writing task in Set 1 was a little more difficult than the independent writing task. That means that examinees were awarded with lower scores on the integrated writing task and received higher scores on the independent writing. Although these differences in task difficulty were statistically significant ($p < .001$), they were not substantial.

Figure 4-3 *Wright Map for the integrated writing task in Set 1*

```
+-------------------------------------------------------------+
|Measr|+Examinee |-Rater                |-Task         |Scale|
|-----+----------+----------------------+--------------+-----|
|  10 + ****.    +                      +              + (5) |
|     |          |                      |              |     |
|   9 + .        +                      +              +     |
|     | *        |                      |              |     |
|   8 + *        +                      +              + --- |
|     | **       |                      |              |     |
|   7 + *        +                      +              +     |
|     | **.      |                      |              |     |
|   6 + **.      +                      +              +  4  |
|     | ****.    |                      |              |     |
|   5 + *****.   +                      +              +     |
|     | **.      |                      |              | --- |
|   4 + ***      +                      +              +     |
|     | ****     |                      |              |     |
|   3 + ******   + 901                  +              +  3  |
|     | **.      |                      |              |     |
|   2 + ***.     +                      +              +     |
|     | *******  | 125                  |              |     |
|   1 + ****.    + 331  487             +              + --- |
|     | *****    | 322  902             |              |     |
*   0 * ****.    * 326  565             * Integrated   *     *
|     | ******.  | 321  346  520  903   |              |     |
|  -1 + ***.     +                      +              +  2  |
|     | ****     |                      |              |     |
|  -2 + *.       + 583                  +              +     |
|     | *******. | 179  222             |              |     |
|  -3 + **       +                      +              +     |
|     | .        |                      |              | --- |
|  -4 + **.      +                      +              +     |
|     | *.       |                      |              |     |
|  -5 + **.      +                      +              +     |
|     | *.       |                      |              |     |
|  -6 + ***.     +                      +              +     |
|     | **.      |                      |              |  1  |
|  -7 + *        +                      +              +     |
|     | *.       |                      |              |     |
|  -8 + *.       +                      +              +     |
|     | .        |                      |              |     |
|  -9 +          +                      +              +     |
|     | .        |                      |              |     |
| -10 + *        +                      +              + --- |
|     |          |                      |              |     |
| -11 +          +                      +              +     |
|     | .        |                      |              |     |
| -12 + *****    +                      +              + (0) |
|-----+----------+----------------------+--------------+-----|
|Measr| * = 2    |-Rater                |-Task         |Scale|
+-------------------------------------------------------------+
```

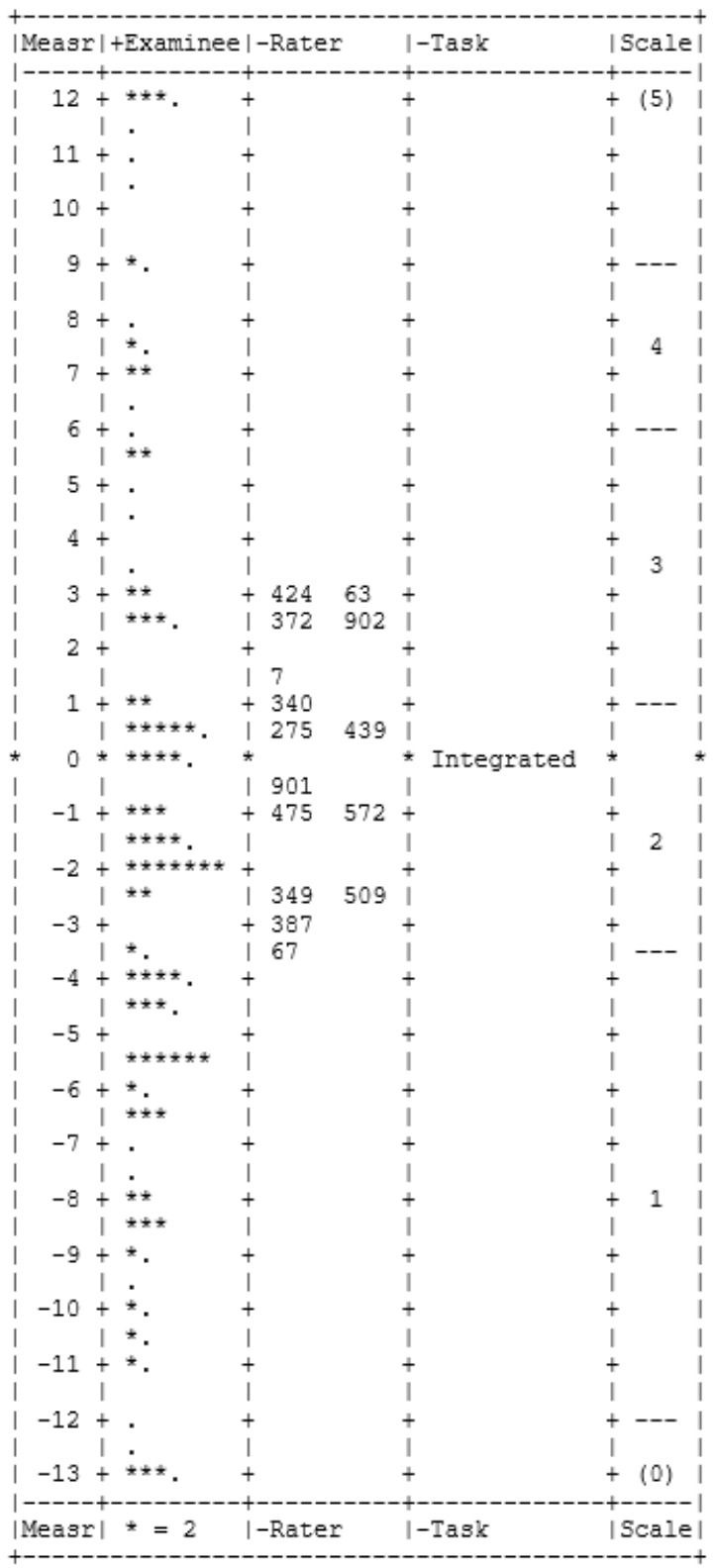
Set 2

The Wright Map of Set 2 (Figure 4-4) reveals a slightly different picture. As in Set 1, proficiency estimate measures for examinees in Set 2 varied substantially, ranging from around above 12 logits (highly proficient examinees) to below -12 logits (examinees with low proficiency). This equals a logit spread of 24 logits, which is about two logits more compared to Set 1. The comparison with the Wright Map for Set 1 also shows that more examinees had lower proficiency estimates in Set 2, while there were less examinees with higher proficiency.

The heterogeneity of raters in terms of their severity looks similar to Set 1 but is even more pronounced in Set 2. The severity measures spread from around three logits for the most severe raters (raters 424 and 63) to the most lenient rater (rater 67) who yielded a severity measure of over 3.5 logits, which equals a logit spread of six-and-a-half logits and is even higher compared to Set 1. There was also a higher number of raters who could be considered severe or lenient in Set 2 with their severity estimates being ≥ 1.0 or ≤-1.0 logits.

Insights on task difficulty can again best be provided by the joint analysis of independent and integrated writing task. As the Wright Map for Set 2 in Appendix H reveals, the integrated writing task in Set 2 had higher difficulty estimates compared to the independent writing, but in Set 2 the logit spread (1.96) was more than twice as high as in Set 1. That is, the MFRM not only revealed a statistically significant ($p < .001$) but also a substantial difference in task difficulty within Set 2 and across sets.

Figure 4-4 *Wright Map for the integrated writing task in Set 2*



4.4.1.2 Separation statistics

The examinee measurement report, and here especially the separation statistics, inform about whether different levels of integrated writing proficiency estimates can reliably be distinguished (RQ 1a).

As Table 4.2 shows, the separation statistics for both sets were similar. The separation index, or examinee strata, provides information about the number of different proficiency levels that could be measured. The values of 4.39 (for Set 1) and 3.92 (for Set 2) for the separation index, as well as the high reliability values of above .80 indicated that there were around four statistically different levels of examinees proficiency which could be reliably distinguished (Eckes, 2015a).

Table 4.2 *Separation statistics*

	Set 1	**Set 2**
Separation (strata) index	4.39	3.92
Separation reliability	.90	.88

4.4.1.3 Rating Scale functioning

The quality of the rating scale is assessed by scale functioning analysis. Amongst others, the scale statistics inform about the consistent use of the rating scale, and whether raters make use of all parts of the scale or whether there is evidence for a restricted range (see Barkaoi, 2014b).

The results of the FACETS output for the integrated writing scale for the two different sets is displayed in Table 4.3. The first column displays the six levels of the holistic rating scale from 0-5 points, the second and third column show the frequency and the percentage a

given score was assigned across all raters and all integrated writing performances. The average test-taker ability for each score level is displayed in the next column. Since examinees with higher ability are assigned with higher scores, these measures are expected to increase from scale level 0 to 5. Column 5 shows the predicted ability measure by the model, and the outfit mean square (OMS) is reported in the last column. A score of 1.0 for the OMS indicates that the observed and the expected measures are equal.

Table 4.3 Scale statistics

Scale levels	Observed counts		*Average measure*	*Expected measure*	*OMS*
	Freq.	%			
Set 1					
0	7	2	-9.69	-9.48	.8
1	86	20	-5.42	-5.48	1.1
2	132	30	-.70	-.68	1.0
3	125	28	2.38	2.52	.9
4	65	15	5.88	5.57	.8
5	39	6	7.66	7.82	1.1
Set 2					
0	7	2	-11.57	-11.84	.8
1	116	37	-6.63	-6.59	1.0
2	100	32	-1.44	-1.43	.9
3	59	19	2.56	2.58	.7
4	24	8	7.40	6.95	.7
5	10	3	8.45	8.85	1.5

As can be seen in Table 4.3, only a small number of integrated writing performances were assigned scores at the ends of the scale, i.e. 0 and 5 points. In general, raters scored a high number of examinees responses with only two or three points on the rating scale. 80% of the performances were placed at scale level 3 and below in Set 1, in Set 2 this percentage was even higher with 90%.

The observed average measures show that the rating scale functioned as expected since higher scores on the scale are associated with higher measures. The expected scores by the measurement model is similar to the observed average measure, hence reported the OMS indices were equal or very close to 1.

4.4.1.4 Rater measurement report

The rater measurement reports will shed light on the issue of reliability and consistency of rating the integrated writing performances (RQ 1b).

Set 1

As already evident from the Wright Map (Figure 4-3), Table 4.4 shows that for Set 1 rater 901 was the most severe rater with a severity measure of 3.24 logits, while rater 222 was the most lenient rater with a measure of -.2.58 logits.

The high logit spread of 5.82 suggests a heterogeneous group of raters. The separation index (strata) and the separation reliability provide further evidence that the group of raters was rather heterogeneous: The calculated value of 5.13 for the separation index suggests that there were more than five statistically distinct groups among the 15 raters. Rater separation reliability was .93, confirming the heterogeneity of severity measures since a high value for rater separation reliability, i.e. close to 1, is not the desired goal – in contrast to the examinee separation reliability (Eckes, 2015a).

The average scores a rater assigned are shown in the column "Observed average". Differences in these scores, for example if the observed average of a rater is higher than other raters' observed scores, could either be related to a rater's severity or the high

proficiency of an examinee that was assigned to him/her. The column "Fair average" therefore provides further insights into this problem since it "adjusts the observed average for the average level of proficiency in the rater's sample of test takers" (Eckes, 2019, p. 164). For example, the observed average for Rater 901 was 2.05, the fair average was 1.66. This difference indicates that Rater 901 rated performances with a high proficiency level. Rater 487 had almost the exact same observed average (2.05), but the fair average was much higher (2.12), indicating that Rater 487 rated performances from examinees with a lower proficiency level.

Table 4.4 *Rater measurement report for the integrated writing task of Set 1*

Rater	Severity Measure	*SE*	MS_W	t_W	MS_U	t_U	Observed average	Fair average	*N* of ratings
901	3.24	.49	1.04	.2	.92	.0	2.05	1.66	40
125	1.71	.48	1.08	.3	1.03	.1	2.45	1.98	60
331	1.21	.32	.96	-.1	.86	-.3	2.29	2.08	40
487	1.01	.43	.96	.0	.79	.0	2.00	2.12	90
322	.63	.39	.86	-.4	.99	-1.3	2.10	2.20	120
902	.56	.49	.56	-1.5	.44	.0	2.20	2.21	60
326	.19	.38	.80	-.7	1.00	-.8	2.30	2.30	40
565	.15	.32	.87	-.4	.77	-.6	2.47	2.31	60
321	-.27	.39	.90	-.2	.75	.0	2.53	2.42	60
346	-.50	.37	.70	.0	.97	-.1	2.53	2.49	60
903	-.50	.45	.93	-.1	.90	-.3	2.40	2.49	40
520	-.51	.28	.99	.0	.91	1.0	2.57	2.49	90
583	-2.0	.43	.95	.0	1.35	1.4	2.7	2.90	48
179	-2.33	.32	1.40	1.6	1.36	-.1	3.33	2.98	80
222	-2.58	.40	.92	-.1	.92		2.87	3.04	60

Note. MS_W = mean-square infit statistic. t_W = standardized infit statistic. MS_U = mean-square outfit statistic. t_U = standardized outfit statistic.

The infit and outfit statistics provide information on rater consistency, i.e. "the degree to which a rater is internally self-

consistent across test takers, criteria, and task and is able to implement the rating scale to make distinctions among test takers' performances" (Barkaoui, 2014b, p. 1308). In judging rater fit, infit statistics (short for "information weighted fit statistics") are considered more important, since infit is more sensitive to unexpected ratings (Eckes, 2015a). Both fit statistics are expected to have a value of 1.0. When raters do have more variation in their ratings than expected by the model, this *misfit* will be indicated by fit values greater than 1.0. On the contrary, values below 1.0 indicate an *overfit*, meaning that ratings showed much less variation than expected. Misfit is generally considered as more problematic and as a threat to score interpretations (Eckes, 2015a).

Fit statistics with values between 0.5 and 1.5 are considered acceptable (Linacre, 2021; Barkaoui, 2014b; Eckes, 2015a). According to Linacre (2021), rating patterns with fit values above 1.5. are "noisy", i.e. ratings are more erratic, while fit values below 0.5 indicate "muted" ratings, meaning that there is too little variation.

Looking at the rater measurement report for Set 1, one can see that almost all fit statistics were within the acceptable range. Only the fit values for Rater 902 showed a relatively high overfit, the mean-square outfit statistic (.44) was even a little outside the acceptable range of .50. The ratings of Rater 902 could be classified as "muted", showing too little variation in the ratings.

Set 2

Looking at the rater severity measures for the integrated writing task in Set 2 (Table 4.5), it is obvious that rater severity varied to a great extent. Rater 63 was the most severe rater with a severity

measure of 2.97 logits, while rater 67 was the most lenient rater with a measure of -.3.56 logits.

As in Set 1, proof of rater heterogeneity was provided by additional data. The rater separation index (strata) had a value of 5.62, separation reliability was .94. This indicates that among the 15 raters, there are five groups that could be distinguished in terms of severity.

Table 4.5 *Rater measurement report for the integrated writing task of Set 2*

Rater	Severity Measure	*SE*	MS_W	t_W	MS_U	t_U	Observed average	Fair average	*N* of ratings
63	2.97	.57	.64	-8	.55	-.5	1.70	1.21	40
424	2.92	.58	1.07	.3	1.01	.1	2.10	1.21	40
902	2.58	.57	1.09	.3	.71	.0	1.45	1.28	40
372	2.28	.63	.90	-.1	.70	.0	1.15	1.34	60
7	1.50	.48	.69	-.9	.52	-.9	1.73	1.54	40
340	.97	.49	1.13	.5	1.10	.3	2.30	1.67	40
275	.47	.45	.94	.0	1.30	.7	2.30	1.79	40
439	.27	.54	1.38	1.1	1.30	.7	2.20	1.84	40
901	-.39	.55	.76	-.4	.63	-.6	2.15	1.96	40
572	-.99	.48	1.17	.6	1.02	.1	2.50	2.07	48
475	-1.18	.40	.79	-.7	.70	-.6	2.50	2.10	40
509	-2.32	.50	1.28	.9	1.17	.4	1.79	2.36	60
349	-2.59	.56	1.12	.4	1.67	1.0	2.90	2.43	60
387	-2.92	.47	.53	-1.5	.38	-1.4	2.00	2.52	60
67	-3.56	.52	1.04	.2	.82	.0	2.05	2.68	40

Note. MS_W = mean-square infit statistic. t_W = standardized infit statistic. MS_U = mean-square outfit statistic. t_U = standardized outfit statistic.

For the observed average, i.e. the mean rating of a rater across all examinees, it is striking that on average raters assigned lower scores to integrated writing performances in Set 2 (*M*=2.06) compared to Set 1 (*M*=2.46). While in Set 1, the values for observed average ranged from 1.66 to 3.33, in Set 2 the minimum observed average was 1.15 (Rater 372), and the highest observed average was

2.90 (Rater 349). Again, a closer look at the fair averages provide further insights whether differences in scores were related to rater severity or proficiency of examinees. For example, Rater 424 and Rater 67 had almost identical measures for observed average (2.10 and 2.05 respectively). The fair average for Rater 424, though, was much lower (1.21), indicating that on average the proficiency level of examinees rated by Rater 424 was high. By contrast, the fair average for Rater 67 was much higher than his/her observed average (2.68), showing that the performances that he/she rated, had a relatively low average level of proficiency.

The fit statistics for Set 2 also confirm that in general raters were able to apply the rating scale consistently. Again, only one rater (Rater 387) showed a high overfit, with a value of .38 for the mean-square outfit statistic being outside the acceptable range.

Overall, the fit statistics from the rater measurement report for Set 1 and Set 2 showed a satisfactorily degree of intra-rater consistency, i.e. the raters applied the rating scale consistently across test-taker performances in both sets.

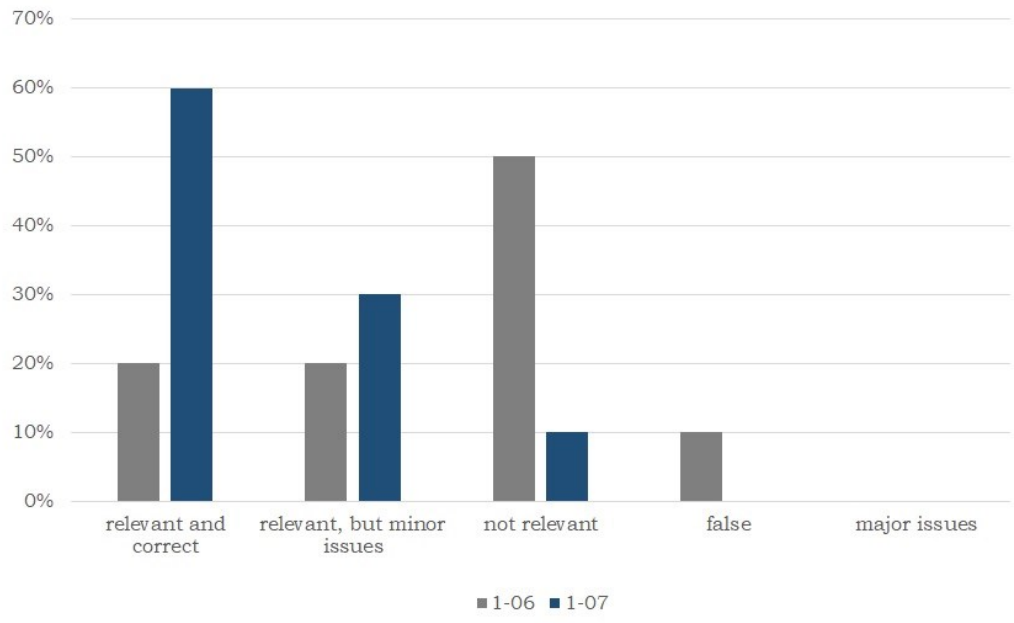
4.4.1.5 Text quality and integrated writing scores

The results from the non-parametric Mann-Whitney-U tests showed statistically significant differences between low- and high-scoring participants only regarding the information that was not comprehensible due to major language problems (see Appendix I).

Nonetheless, the written responses of low- and high-scoring participants varied to some extent, as a closer look at the summaries from two writers revealed. Drawing on findings from Strand 2 (see section 3.4.1), the written performances of one high-scoring writer (participant 1-07) and one low-scoring writer (participant 1-06) were compared.

Looking at results for the relevance and accuracy of information (Figure 4-5), the comparison of the two participants showed, that the proportion of relevant and correct information for participant 1-07 (high-scorer) was much higher (60%) compared to the proportion for the low-scoring participant 1-06 (20%). The low-scoring writer also included much more irrelevant information in his response. 50% of information in his summary was not relevant in relation to the task requirements, compared to only 10% for the high-scoring participant. The high-scorer also did not include false information in his written response. Both summaries did not contain any information that was not comprehensible due to major language issues.

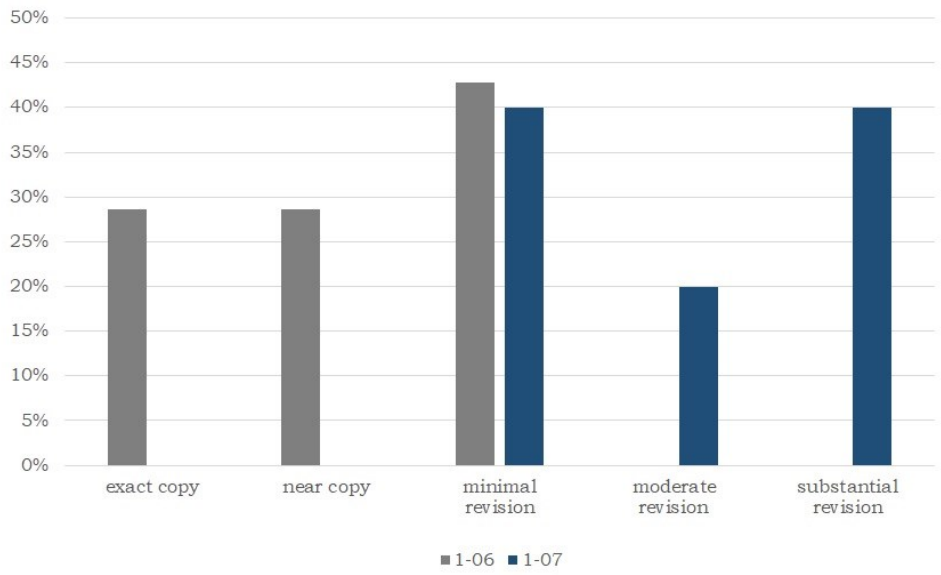
Figure 4-5 *Comparison low- vs. high-scoring participant: Relevance and accuracy of information*



The way both participants transformed the language of the input material also revealed differences (see Figure 4-6). The integrated

writing performance of participant 1-06 showed a much higher proximity to the source material. Around 28% of his text were directly copied from the sources, another 28% were defined as near copies, the rest was only minimally revised. On the contrary, participant 1-07 transformed the language of the input material to a much greater extent. Even though the proportion of minimally revised information was almost as high as for the low-scoring participant, his text also included 20% of moderate, as well as 40% of substantial revisions. The high-scorer did not copy directly from the sources.

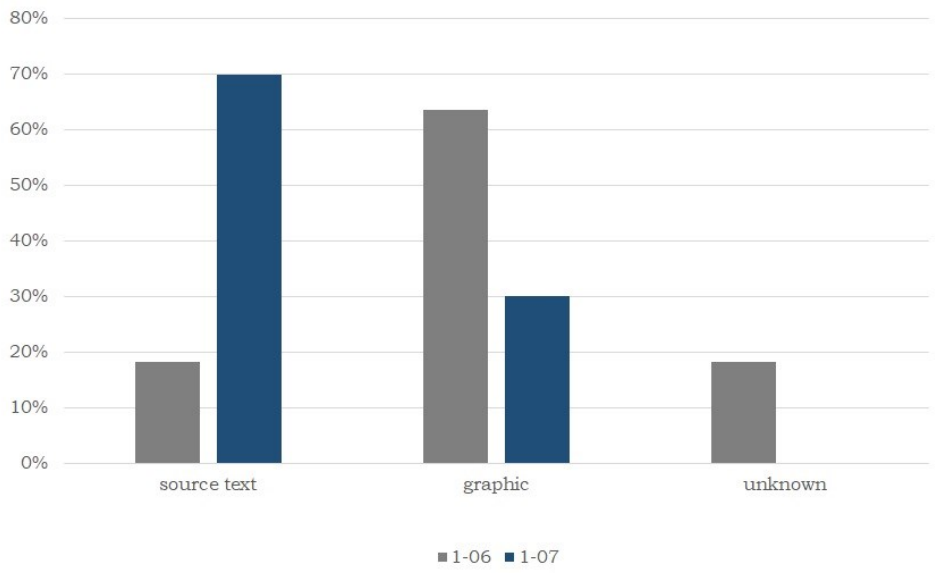
Figure 4-6 *Comparison low- vs. high-scoring participant: Paraphrase type*



The performances also varied regarding the use of the two different sources (Figure 4-7). In his summary, participant 1-06 mostly drew on the graphical input, only 18% of the information were taken from the source text. He also included some information that was not based on the input material. On the contrary, the high-scoring participant 1-07 based his written response mainly on information taken from the source text (70%). He only included a

small amount of information from the graphical input (30%) and added no additional information.

Figure 4-7 *Comparison low- vs. high-scoring participant: Origin of information*



Even though the group comparison did not reveal statistically significant differences between low- and high-scoring participants, an exemplification of two writers showed that the performances of a successful and a less successful writer still varied in terms of content and integration style.

4.4.2 Reading-writing relations in rating integrated writing performances

To explore the relationship between independent and integrated writing performances, as well as the relation between integrated writing performance and reading comprehension (RQ 2), a correlation analysis was conducted. Integrated writing scores in terms of fair average measures were correlated with scores that participants yielded in the independent writing task and scores they

received in the reading component of the digital TestDaF. The analysis included all participants from the piloting stage. Results across sets, and for each set separately, are displayed in Table 4.6.

Table 4.6 *Pearson correlation analysis results*

		Independent writing scores	**Reading scores**
Integrated writing scores	overall	.593**	.525**
	Set 1	.630**	.545**
	Set 2	.557**	.501**

Note. ** p < .01

As can be seen, there were statistically significant high correlations between the scores on both writing tasks (r = .593, p < .01). The shared variance of 35% across the two sets ($.593^2$*100) indicated some overlap between independent and integrated writing performances.

The integrated writing scores also correlated strongly with the reading scores (r = .525, p < .01). As expected, reading and writing ability are somewhat related in integrated writing, the shared variance between scores on the integrated writing task and the scores on the reading component of the digital TestDaF showed that 27% of the integrated writing scores could be explained by reading ability.

4.5 Discussion

Analyzing the rating data from piloting two task versions of the integrated writing task included in the digital TestDaF revealed the following in relation to the different facets:

- Examinees: Examinees' writing proficiency varied to a great extent; separation statistics showed that four different levels of integrated writing proficiency estimates can reliably be distinguished.

It might seem critical that the number of reliably distinct levels of examinee proficiency by the measurement system was lower than the supposed six levels in the rating scale of the digital TestDaF. This could be explained by the fact that there were hardly no examinees at the highest and the lowest level of proficiency in both sets. Consequently, the spread of examinees proficiency was limited to less proficiency levels than expected. In addition, the writing scores are not based on the integrated writing performance alone, but rather derived from the ratings of the integrated and the independent writing performances. Given the fact, that the aim of the TestDaF is to relate examinee's performance in each of the four test components to one of the three TestDaF proficiency levels (see 1.2), the examinee strata for the integrated writing performance seems satisfying (Barkaoui, 2014b; Bond & Fox, 2007)

- Rater facet: Raters differed in terms of their severity/leniency when they rated written performances, but apart from very few exemptions, their ratings were consistent.

The overfit of Rater 902 in Set 1 and Rater 387 in Set could either be related to a rater effect like central tendency, i.e. raters prefer the categories in the middle of the scale and avoid the extreme categories at the ends of the scale. The overfit might have also been caused by a very homogenous group of examinees that were assigned to these

raters. Possible causes for the overfit could be identified by two measures: a) a rater-related partial credit MFRM, and b) a qualitative analysis of the written performances and the accompanying justifications of scores. Taking into account that the holistic rating scales were applied for the first time at this piloting stage and rater reliability was overall satisfying, this issue was not further investigated.

- Task: The correspondence of examinee proficiency, rater severity and task difficulty on the logit scale, showed that the task was not too difficult or too easy for the specific population of test takers in Set 1, but a little too difficult for the population in Set 2. The integrated writing task in both sets was more difficult compared to the independent writing task, in Set 2 the difference was substantial.
- Rating scale functioning showed that raters used the whole range of the scale, but most scores were assigned to levels 2 and 3, only a few performances were placed at the ends of the scale, i.e. assigned with scores of 0 or 5 points.

Overall, the quantitative analysis of rating data confirmed the validity of the assessment procedures for the integrated writing task of the digital TestDaF.

The linking of product and scoring data did not confirm the assumption, that the quality of the written performances was reflected in the score. In contrast to the qualitative analysis of low- and high-proficient learners in Chapter 2, the comparison of low- and high-scoring participants in this strand did not reveal statistically significant differences between the two groups, except for information that was incomprehensible due to language problems. On the individual level, differences existed, as the comparison of two participants revealed.

This unexpected outcome may be related to different aspects: Firstly, the participants were unevenly distributed among the two groups, with only six participants in the high-scoring group. The Mann-Whitney-U test usually works with unequal sample sizes, but the statistical power is low for small sample sizes, and diminishes with uneven distributed groups. In addition, in Strand 2, participants were divided into two groups according to their overall language proficiency as measured by a C-test (see 3.4.2) In this strand (Strand 3), participants were assigned to one of two groups based on their writing score, i.e. the fair average they yielded in the integrated writing task. The cut-score was set at 2.50 logits which equals the midpoint region of the 6-point rating scale. This means, that the two groups included distinct levels of proficiency and covered more than one level of the rating scale each. Especially problematic are performances around the set cut-score. A participant with a proficiency estimate of just above 2.50 logits was assigned to the high-scoring group, while another participant just below that threshold was assigned to the low-scoring group. Their writing performances might therefore be similar, but the assignment to two distinct groups might have had an impact on the analysis outcome. And finally, the quality of the integrated writing performances was measured by looking at source use and integration style only, which implies that other characteristics like organization or accuracy as well impact the rating of integrated writing performances which have not been addressed in this study.

Regarding the construct underlying the integrated writing task, this study found high correlations with both the independent writing and the reading comprehension test, confirming results by Shin and Ewert (2015), implying that both reading and writing ability are involved in integrated writing assessment. However, the shared variance of 35% between the two writing tasks also implies that they

both tap into the same ability, i.e. writing, but still measure different dimensions of the writing construct. The inclusion of two different tasks, i.e. one independent and one integrated, in the writing component of the digital TestDaF therefore broadens the construct being measured. A composite score (Gebril, 2010) could then be regarded as a valid and reliable measure of writing proficiency.

5 CONCLUSION

The overall research aim of this dissertation was to provide evidence for the claim that the integrated writing task of the digital TestDaF is a valid and reliable measure for academic writing in the context of university admission in Germany. Applying a mixed-method design, empirical evidence for backing the following assumptions was collected:

- Assumption 1: The integrated writing task of the digital TestDaF elicits cognitive processes that are typical for writing from sources within the context of academic writing.
- Assumption 2: The (successful) processing and transformation of the input material is evident in the written product.
- Assumption 3: The quality of the written summary is reflected in the score.

Assumption 1 was addressed in the first research strand. Using a combination of eye-tracking and stimulated recalls, cognitive processes of international study applicants during task completion were investigated. Findings from the quantitative and qualitative data revealed, that test takers engaged in a variety of cognitive processes related to basic processes of reading and writing, but also employed processes that integrated reading and writing in so-called shared processes. The identified processes confirmed findings from previous research and allow for linking them to existing L2 integrated writing models like *discourse synthesis*. These models are postulated representations of cognitive processes involved in 'real' source-based writing since they are often based on empirical evidence. Therefore, the findings from this research strand support the extrapolation inference for the postulated interpretive argument (see section 1.4) because the integrated writing task involves the same processes as similar tasks in the TLU domain (Kane, 2013a).

In Strand 2, integrated writing performances were analyzed. The analysis focused on both the incorporation of relevant information from the sources and the use of paraphrases to transform the language of the input material. Results not only showed differences between participants at distinct levels of proficiency (low vs. high), but also revealed that the processing of the written source text was distinct from the processing of the graphical input with respect to source use and integration style. Since the coding scheme used in this analysis was based on the rating scale in use for the integrated writing task, the results from this strand can be used as a backing for an explanation inference, i.e. that the expected scores are attributed to a construct of language proficiency (Knoch & Chapelle, 2018; Yang, 2014). This inference is not an essential part of Kane's *IUA* (see1.4), but can be included to add to a theory-defined construct of language ability, and therefore improve the interpretation and use of test scores.

The last research strand used quantitative methods to look into the reliability of rating integrated writing performances. Results from MFRM confirmed in relation to the rater facet that raters were able to identify different levels of examinee's writing proficiency, and that they applied the scale consistently. In relation to scale functioning, the results showed that the scale was able to distinguish among levels of proficiency, i.e. test takers were placed at a certain level based on characteristics of their performance (Kane, 2013a). Because the scale functioned as intended, and rater rated reliably, the empirical evidence gathered in this strand allows for backing the evaluation/scoring inference of the interpretive argument (Knoch & Chapelle, 2018).

To conclude, by taking a closer look at the construct underlying integrated writing from three different perspectives, i.e. processes, products, and scoring, the thesis was able to gather comprehensive

empirical evidence to support a validity argument for the integrated writing task of the digital TestDaF.

6 IMPLICATIONS

The findings of this study have implications for language learning and teaching, as well as for rater training and monitoring. The results can also help improving rating scale design.

Language learning and teaching

The eye-tracking data and the stimulated recalls provided valuable insights into the cognitive processes of language learners who were preparing for taking up studies in Germany. The integrated writing task elicited a variety of reading, writing and shared processes, and required test-taking strategies. It became evident, that reading was an integral part of the writing process. The comprehension of the instructions and the source material proved to be crucial for successfully working on the task. The existence of shared processes showed that reading processes were intertwined with writing processes, proving specific reading-writing-relations in integrated writing.

Especially the stimulated recalls revealed difficulties participants had with the newly introduced task format. To some extent these problems were related to the fact, that at the time of data collection almost no preparation material for the digital TestDaF was available. Participants had access to a model test, but no further explanation of the task requirements were provided. They also did not receive any feedback on their answers in the model test. But the difficulties were also caused by lack of familiarity with source-based writing. Participants almost had no prior experience with summary writing, and therefore lacked the necessary strategies to cope with the task demands.

These issues demonstrate the need for a specific approach to preparatory language classes. Since integrated test tasks require the combination of different skills like reading and writing, the focus of test preparation should shift from learning and teaching skills separately to fostering a learning-oriented approach to test preparation (Green, 2017) with a focus on competencies across skills – which is also relevant for dealing with communicative tasks in the TLU domain (see Grabe & Zhang, 2013).

This approach was adopted for a comprehensive theoretical concept for test preparation that was developed at the TestDaF-Institut (Kecker, Kleppin, & Zimmermann, 2019a). The concept is based on two central aspects: a) focusing on competences underlying the test tasks across skills and across test components, and b) raising test takers' awareness for the requirements of the test tasks and how these are related to the TLU domain. This would require language learners and teachers to identify the competences underlying the single test tasks and to recognize how they are linked to other tasks in other components of the digital TestDaF. For example, in order to successfully work on the integrated writing task, test takers would need to identify differences and commonalities in the input material, recognize causal relationships within and across the sources and link this to the given question in the instruction, as well as to verbalize information from a graphical input. The latter is also needed for a task in the speaking component, where test takers are presented a graphical input, listen to a short statement and then have to present their own point of view by taking into account the information from the graphical input. In this sense, practicing this specific competence does not only prepare test takers for one single writing task, but helps them in preparing for other test tasks.

In addition, the new approach for test preparation fosters activities that sensitizes test takers for the relation of the competences underlying the test tasks and the requirements of communicative tasks at university. The authenticity of the test task (Bachman, 1990; Bachman & Palmer, 1996) may have also been enhanced by the digitalization of the test format (Suvorov & Hegelheimer, 2014). Computer literacy as well as the processing of different media, e.g. audios, videos, visuals, is a prerequisite for academic success. Hence, the incorporation of the computer in the language classroom or opportunities for digital language learning can be regarded as a positive washback effect (Kecker, Depner, Marks, Schwarz, & Zimmermann, 2019b) – if test takers are enabled to develop strategies to cope with reading, listening, writing, and speaking online (see Hirvela, 2005)

Although test takers approached the integrated writing task following a similar pattern (see section 2.5), the analysis of the writing process revealed a great variation between participants. Test preparation therefore should not be a "one-fits-all"-approach, rather taking into account individual preferences and prior knowledge of the test takers (Kecker et al., 2019a).

Rater training and monitoring

The high reliability of ratings proved that the intensive rater training was successful. Raters could apply the rating scale as intended and the scale design seemed appropriate to capture differences of the integrated writing performances at various levels of proficiency.

The accompanying calibration material (see section 4.3.3) surely played an important role in establishing a common understanding of the task requirements across raters. Especially with regard to

content raters need to know what relevant information from the sources is expected to be included in the responses of test takers. To date, there is no data available to what extent raters familiarize themselves with the material before starting rating. For further rater monitoring, it could therefore be worth considering to give raters only access to their assigned ratings in the online tool if they have worked through the calibration material, i.e. they have read the description of what to expect with regard to content and rated the accompanying benchmark performances. In cases where their rating differs from the benchmark, they are provided with some sort of feedback. Even though this approach would result in a higher workload for raters (and testing experts at TestDaF-Institut), it might enhance the reliability of rating integrated writing performances.

When rating integrated writing performances, raters often focus on the amount of overlap between the input material and the written texts. The approach adopted in this study (Keck, 2006; see section 3.3.5.1) might be helpful in rater training. Looking more closely at *what* is being copied verbatim – keywords vs. general phrases – could be more helpful for raters to decide about the linguistic range of test takers than simply looking at the similarity between source material and test takers' responses. In addition, the in-depth analysis of the written responses showed that the information from the source text was transformed differently than the information from the graphical input. Raters also have to be sensitized for this issue.

Rating scale design

Findings of the study implied that the rating scale for the integrated writing task of the digital TestDaF is able to capture salient features of writing performances at different levels of proficiency. Nonetheless, it is also essential to regularly monitor and

modify existing rating scales to further ensure and improve the reliability of test scores (Banerjee, Yan, Chapman, & Elliot, 2015).

Rating scale descriptors are usually 'generic' so that they can be applied to different task versions (Chan et al., 2015). As the analysis of the written performances in this study showed, there are specific characteristics regarding the linguistic transformation of two different types of sources, i.e. written and graphical input material. The results reported in Strand 2 also confirmed findings from previous studies which showed that summary writing is characterized by less restructuring of the input material (Spivey, 1990), and that writers borrow significantly more words from the sources compared to other types of source-based writing (Shi, 2004). These issues are not addressed in the rating scale currently in use. For example, the generic wording of the descriptor for a 5-point performance only refers to the "use of a broad range of complex language to summarize information in one's own words". To confirm the preliminary insights into the specific characteristics of TestDaF integrated writing performances regarding integration style, it is recommended to apply the analysis used in this study to a larger sample of written responses from actual live tests and thereby help to inform a possible revision of the writing scale.

Future research

Interpretability of the findings could be further improved by several means: First, including more high-proficient writers in the data collection, i.e. not relying on convenient sampling, but rather carefully select participants. Furthermore by adopting a research design that in addition to eye-tracking data includes keystroke logging for further data triangulation (Michel et al., 2020; Révész et al., 2017). The logging of keystrokes would allow for a more

comprehensive understanding of pausing behavior and revision in the writing process (Chan, 2017; Leijten & van Waes, 2013).

In what way the processing of written input material differs from the processing of graphics has not been in the focus of integrated writing research, yet. The findings of the current study only provided preliminary insights into this specific issue. For a fundamental theory or model of writing from sources in the L2 that has been demanded (e.g. by Cumming, 2013; Knoch & Sitajalabhorn, 2013;), a future research agenda should also take into account integrated writing tasks that are not only text-based. As an alternative to the combination of eye-tracking and stimulated recalls, the use of TAPs might provide more detailed insights into the differences in processing textual and graphical sources in integrated writing tasks.

While the analysis and coding of test takers' responses was done manually in this study, a way forward could be to use part of the analysis as a starting point to build an assisted or automated scoring tool for rating integrated writing performances. Some tools for exploring the writing quality of summary writing like CRAT[42] (Crossley, Kyle, Davenport, & McNamara, 2016) already exist. Unfortunately, they only work for English, and again, the indices only include source text/summary text overlap – without taking into account graphical input material. To consider additional sources besides written texts, and to rethink the quality of summary writing beyond simple overlap could therefore also be of interest for Natural language processing (NLP) researchers.

[42] CRAT is freely available for download. https://www.linguisticanalysistools.org/crat.html

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arras, U. (2012). Im Rahmen eines Hochschulstudiums in Deutschland erforderliche sprachliche Kompetenzen: Ergebnisse einer empirischen Bedarfsanalyse. In T. Tinnefeld (Ed.), *Hochschulischer Fremdsprachenunterricht: Anforderungen – Ausrichtung – Spezifik* (pp. 137–148). Saarbrücken: htw saar.

Asención Delaney, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7(3)*, 140–150. https://doi.org/10.1016/j.jeap.2008.04.001

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. F. (1997). Generalizability theory. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education. Vol 7: Language testing and assessment* (pp. 255–262). Dordrecht, Boston, London: Kluwer Academic Publishing.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bachman, L. F. (2005). Building and Supporting a Case for Test Use. *Language Assessment Quarterly, 2(1)*, 1–34. https://doi.org/10.1207/s15434311laq0201_1

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

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5–19.

Barkaoui, K. (2014a). Examining the impact of L2 proficiency and keyboarding skills on scores on TOEFL-iBT writing tasks. *Language Testing, 31(2)*, 241–259. https://doi.org/10.1177/0265532213509810

Barkaoui, K. (2014b). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment: Volume III: Evaluation, methodology, and interdisciplinary themes* (pp. 1301–1322). Hoboken, NJ, USA: Wiley-Blackwell.

Barkaoui, K. (2015). *Test takers' writing activities during the TOEFL iBT ® writing tasks: A stimulated recall study* (TOEFL iBT Research Report Series No. 25).

Barkaoui, K. (2016). *Examining the cognitive processes engaged by Aptis writing task 4 on paper and on the computer* (ARAGs Research Reports Online No. AR-G/2016/1).

Barkaoui, K., & Knouzi, I. (2018). The effects of writing mode and computer ability on L2 test-takers' essay characteristics and scores. *Assessing Writing, 36*, 19–31. https://doi.org/10.1016/j.asw.2018.02.005

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing, 30(4)*, 441–465. https://doi.org/10.1177/0265532212473244

Bax, S., & Chan, S. (2016). *Researching the cognitive validity of GEPT High-intermediate and Advanced reading: An eye tracking and stimulated recall Study* (LTTC-GEPT Research Reports No. RG-07). Taipeh, Taiwan.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50(1)*, 110–114. https://doi.org/10.1111/jedm.12006

Brennan, R. L. (2001). Generalizability theory. New York: Springer.

Brinkschulte, M. (2012). Akademisches Schreiben in der Fremd- und Zweitsprache Deutsch. In K. Draheim, F. Liebetanz, & S. Vogler-Lipp (Eds.), *Key Competences for Higher Education and Employability. Schreiben(d) lernen im Team: Ein Seminarkonzept für innovative Hochschullehre* (pp. 59–81). Wiesbaden: Springer VS. https://doi.org/10.1007/978-3-531-19129-4_5

Brunfaut, T., Harding, L., & Batty, A. O. (2018). Going online: The effect of mode of delivery on performances and perceptions on an English L2 writing test suite. *Assessing Writing, 36*, 3–18. https://doi.org/10.1016/j.asw.2018.02.003

Brunfaut, T., & McCray, T. (2015). *Looking into test-takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study*. British Council: ARAGs Research Reports (AR/2015/001).

Brunfaut, T., & McCray, T. (2018). Investigating the construct measured by banked gap-fill items: Evidence from eye-tracking. *Language Testing, 35(19)*, 51–73. https://doi.org/10.1177/0265532216677105

Chan, S. (2013). *Establishing the validity of reading-into-writing test tasks for the UK academic context*. PhD thesis. University of Bedfordshire, UK.

Chan, S. (2017). Using keystroke logging to understand writers' processes on a reading-into-writing test. *Language Testing in Asia, 7(1)*, 140. https://doi.org/10.1186/s40468-017-0040-5

Chan, S., Bax, S., & Weir, C. (2018). Researching the comparability of paper-based and computer-based delivery in a high-stakes writing test. *Assessing Writing, 36*, 32–48. https://doi.org/10.1016/j.asw.2018.03.008

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing, 26*, 20–37. https://doi.org/10.1016/j.asw.2015.07.004

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple… *Language Testing, 29(1)*, 19–27. https://doi.org/10.1177/0265532211417211

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511733116

Chapelle, C., Enright, M., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment. Vol II: Evaluation, methodology, and interdisciplinary themes* (pp. 1081-1097). Malden, MA: John Wiley.

Chin, S. J. (2009). Investigating summary writing performance of university students in Taiwan. In *Proceedings of the 26th conference of English teaching and learning in the R.O.C.* (pp. 285–301). Taipeh, Taiwan: Crane Publishing Co., Ltd.

Cho, Y., Rijmen, F., & Novák, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT integrated writing tasks. *Language Testing, 30(4)*, 513–534. https://doi.org/10.1177/0265532213478796

Cho, Y., & Choi, I. (2018). Writing from sources: Does audience matter? *Assessing Writing, 37*, 25–38. https://doi.org/10.1016/j.asw.2018.03.004

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1(3)*, 98–101.

Conklin, K., Pellicer-Sánchez, A., & Carroll, G. (2018). *Eye-tracking: A guide for applied linguistics research*. Cambridge: Cambridge University Press.

Conklin, K., & Pellicer-Sánchez, A. (2016). Using eye-tracking in applied linguistics and second language research. *Second Language Research, 32(3)*, 453–467. https://doi.org/10.1177/0267658316637401

Council of Europe (2001). *Common European framework of references for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Creswell, J. W. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches* (3rd ed.). Los Angeles, London: SAGE.

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Los Angeles: SAGE.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Los Angeles, London, New Delhi, Singapore, Washington DC, Melbourne: SAGE.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.) *Educational Measurement* (pp.443-507). Washington DC: American Council on Education.

Crossley, S. A., Kyle, K., Davenport, J., & McNamara, D. S. (2016). Automatic assessment of constructed response data in a chemistry tutor. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th International Educational Data Mining (EDM) Society Conference* (pp. 336-340).

Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10(1)*, 1–8.

Cumming, A. (2014). Assessing integrated skills. In A. J. Kunnan (Ed.), *The companion to language assessment: Volume I: Abilities, contexts, and learners* (pp. 216–229). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla131

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10(1)*, 5–43. https://doi.org/10.1016/j.asw.2005.02.001

Deygers, B. (2017). *Assessing high-stakes assumptions. A longitudinal mixed-methods study of university entrance language tests, and of the policy that relies on them. Doctoral dissertation*. KU Leuven, Belgium.

Dittmann, J., Geneuss, K. A., Nennstiel, C., & Quast, N. A. (2003). Schreibprobleme im Studium: Eine empirische Untersuchung. In K. Ehlich & A. Steets (Eds.), *Wissenschaftlich schreiben: Lehren und lernen* (pp. 155–185). Berlin, New York: W. De Gruyter.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative, and mixed methodologies*. Oxford, New York, Auckland: Oxford University Press.

Dresing, T., & Pehl, T. (2018). *Praxisbuch Interview, Transkription & Analyse: Anleitungen und Regelsysteme für qualitative Forschende* (8. Aufl.). Marburg: Eigenverlag.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197–221.https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2010). Der Online-Einstufungstest Deutsch als Fremdsprache (onDaF): Theoretische Grundlagen, Konstruktion und Validierung. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/The C-Test: Contributions from Current Research* (pp. 126–191). Frankfurt/M.: Peter Lang.

Eckes, T. (2015a). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (2nd ed.). Frankfurt/M.: Peter Lang.

Eckes, T. (2015b). Validität: Flexionen eines polymorphen Konzepts. In J. Böcker & A. Stauch (Eds.), *Fremdsprachen Lebenslang Lernen: Vol. 4.*

Konzepte aus der Sprachlehrforschung – Impulse für die Praxis: Festschrift für Karin Kleppin (pp. 449–468). Frankfurt: Peter Lang.

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated assessment. In V. Arydoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment: Vol. 1. Fundamental techniques* (pp. 153–175). New York, NY: Routledge.

Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2016). Assessing writing. In Tsagari, Dina & Banerjee, Jayanti (Eds.), *Handbook of second language assessment* (pp. 147–164). Boston, Mass.: De Gruyter.

Ehlich, K. (2003). Universitäre Textarten, universitäre Struktur. In K. Ehlich & A. Steets (Eds.), *Wissenschaftlich schreiben: Lehren und lernen* (pp. 13–28). Berlin, New York: W. De Gruyter.

Ehlich, K., & Steets, A. (2003). Wissenschaftliche Schreibanforderungen in den Disziplinen. Eine Umfrage unter ProfessorInnen der LMU. In K. Ehlich & A. Steets (Eds.), *Wissenschaftlich schreiben: Lehren und lernen* (pp. 129–154). Berlin, New York: W. De Gruyter.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? Language Assessment Quarterly, 2(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper.* (TOEFL Monograph Series: MS-17). Princeton, NJ: Educational Testing Service.

Ericsson, K., & Simon, H. (1996). *Protocol analysis. Verbal reports as data* (3rd ed.). Cambridge, MA: MIT Press.

Field, J. (2011). Cognitive Validity. In L. Taylor (Ed.), *Examining speaking* (pp. 65–110). Cambridge: Cambridge University Press.

Field, J. (2013). Cognitive validity. In A. Geranpayeh & L. Taylor (Eds.), *Examining listening* (pp. 77–150). Cambridge: Cambridge University Press.

Flower, L., & Hayes, J. R. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3–30). Lawrence Earlbaum Associates.

Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication, 32(4)*, 365–387.

Freedman, E. G., Shah P. (2002) Toward a Model of Knowledge-Based Graph Comprehension. In M. Hegarty, B. Meyer & N. H. Narayanan. (Eds.) *Diagrammatic Representation and Inference. Diagrams 2002* (pp. 18-30). Berlin, Heidelberg: Springer https://doi.org/10.1007/3-540-46037-3_3

Garner, R., Gillingham, M. G., & White, C. S. (1989). Effects of 'seductive details' on macroprocessing and microprocessing in adults and children. *Cognition and Instruction, 6(1),* 41–57. https://doi.org/10.1207/s1532690xci0601_2

Gass, S. M., & Mackey, A. (2017). *Stimulated recall methodology in applied linguistics and L2 research.* New York, NY: Routledge.

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

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow Working Papers in Second or Foreign Language Assessment,* (7), 47–84.

Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly, 10(1)*, 9–27. https://doi.org/10.1080/15434303.2011.642040

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing, 21*, 56–73.

Gesellschaft für akademische Studienvorbereitung und Testentwicklung (g.a.s.t.) (2020). *Daten kompakt – Compact data*. Bochum: TestDaF-Institut. retrieved from www.testdaf.de.

Ginther, A. (2001). *Effects of the presence and absence of visuals on performance on TOEFL CBT listening-comprehensive stimuli.* ETS Research Report (RR-01-16). http://dx.doi.org/10.1002/j.2333-8504.2001.tb01858.x

Godfroid, A., & Hui, B. (2020). Five common pitfalls in eye-tracking research. *Second Language Research, 36(3)*, 277–305. https://doi.org/10.1177/0267658320921218

Godfroid, A., Winke, P., & Conklin, K. (2020). Exploring the depths of second language processing with eye tracking: An introduction. *Second Language Research, 36(3)*, 243–255. https://doi.org/10.1177/0267658320922578

Grabe, W., & Zhang, C. (2013). Second language reading-writing relations. In Alice S. Horning & Elizabeth W. Kraemer (Ed.), *Reconnecting reading and writing* (pp. 108–133). Anderson, South Carolina, Fort Collins, Colorado: Parlor Press LLC; The WAC Clearinghouse.

Green, A. (2017). Learning-oriented language test preparation materials: A contradiction in terms? *Papers in Language Testing and Assessment, 6(1)*, 112-132.

Grießhammer, E. (2011). *Der Schreibprozess beim wissenschaftlichen Schreiben in der Fremdsprache Deutsch und Möglichkeiten seiner Unterstützung. Beiträge zur Schreibzentrumsforschung: Vol. 3.* Frankfurt/Oder: Europa-Universität Viadrina.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In M. C. Levy & S. Ransdell (Eds.), *The science of*

writing: Theories, methods, individual differences, and applications. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hirvela, A. (2004). *Connecting Reading and Writing in Second Language Writing Instruction.* Ann Arbor, MI: The University of Michigan Press.

Hirvela, A. (2005). Computer-based reading and writing across the curriculum: Two case studies of L2 writers. *Computers and Composition, 22(3)*, 337–356. https://doi.org/10.1016/j.compcom.2005.05.005

Hirvela, A., & Du, Q. (2013). "Why am I paraphrasing? ": Undergraduate ESL writers' engagement with source-based academic writing and reading. *Journal of English for Academic Purposes, 12(2)*, 87–98. https://doi.org/10.1016/j.jeap.2012.11.005

Hochschulrektorenkonferenz, & Kultusministerkonferenz (2015). *Rahmenordnung über Deutsche Sprachprüfungen für das Studium an deutschen Hochschulen (RO-DT): Beschluss der HRK vom 08.06.2004 und der KMK vom 25.06.2004 i.d.F. der HRK vom 10.11.2015 und der KMK vom 12.11.2015.*

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford: Oxford University Press.

Hood, S. (2008). Summary writing in academic contexts: Implicating meaning in processes of change. *Linguistics and Education*, 19(4), 351–365. https://doi.org/10.1016/j.linged.2008.06.003

Howard, R. M. (1993). A plagiarism pentimento. *Journal of Teaching Writing, 11*, 233-246.

Hruschka, D. J., Schwartz, D., St. John, D. C., Picone-Decaro, E., Jenkins, R. A., & Carey, J. W. (2004). Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field Methods, 16(3)*, 307–331. https://doi.org/10.1177/1525822X04266540

Jin, Y., & Yan, M. (2017). Computer literacy and the construct validity of a high-stakes computer-based writing assessment. *Language Assessment Quarterly, 14(2)*, 101–119. https://doi.org/10.1080/15434303.2016.1261293

Kane, M. (2011). Validating score interpretations and uses. *Language Testing, 29(1)*, 3–17. https://doi.org/10.1177/0265532211417210

Kane, M. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50(1)*, 1–73. https://doi.org/10.1111/jedm.12000

Kane, M. (2013b). Articulating a validity argument. In G. Fulcher, & F. Davidson (Eds). *The Routledge handbook of language testing* (pp. 34-47). London, New York: Routledge.

Katz, I. R., Xi, X., Kim, H.-J., & Cheng, P. (2004). Elicited speech from graph items on the Test of Spoken English. *ETS Research Report Series*, 1, i–31. https://doi.org/10.1002/j.2333-8504.2004.tb01933.x

Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15(4)*, 261–278. https://doi.org/10.1016/j.jslw.2006.09.006

Kecker, G. (2010). Validität und Validierung von Sprachprüfungen. In A. Berndt & K. Kleppin (Eds.), *Sprachlehrforschung: Theorie und Empirie: Festschrift für Rüdiger Grotjahn* (pp. 129–146). Frankfurt: Peter Lang.

Kecker, G., Depner, G., Marks, D., Schwarz, L., & Zimmermann, S. (2019a). Die deutsche Sprache weltweit fördern: Was können Sprachprüfungen dazu beitragen? In U. Ammon & G. Schmidt (Eds.), *Förderung der deutschen Sprache weltweit: Vorschläge, Ansätze und Konzepte* (pp. 393–406). Berlin: de Gruyter.

Kecker, G., Kleppin, K., & Zimmermann, S. (2019b). *Konzept zur Vorbereitung auf den digitalen TestDaF*. Unveröffentlichtes Manuskript. Bochum: g.a.s.t.

Kecker, G., Zimmermann, S., & Eckes, T. (in press). Der Weg zum digitalen TestDaF: Konzeption, Entwicklung und Validierung. In P. Gretsch & N. Wulff (Eds.), *Deutsch als Zweit- und Fremdsprache in Schule und Beruf: Eine Festschrift für Gabriele Kniffka* (pp. 393-410). Paderborn: Brill Schöningh.

Kennedy, M. L. (1985). The composing process of college students writing from sources. *Written Communication, 2(4)*, 434–456.

Keseling, G. (1993). *Schreibprozeß und Textstruktur: Empirische Untersuchungen zur Produktion von Zusammenfassungen*. De Gruyter.

Kim, S.-Y., Graham, S. S., Ahn, S., Olson, M. K., Card, D. J., Kessler, M. M., De Vasto, D. M., Roberts, L. R., & Bubacy, F. A. (2016). Correcting biased Cohen's kappa in NVivo. *Communication Methods and Measures, 10(4)*, 217–232. https://doi.org/10.1080/19312458.2016.1227772

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

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction, 7(3)*, 161–195. https://doi.org/10.1207/s1532690xci0703_1

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85(5)*, 363–394. https://doi.org/10.1037//0033-295X.85.5.363

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

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: The effect of background knowledge revisited. *Language Testing, 23(1)*, 99–130. https://doi.org/10.1191/0265532206lt323oa

Kuckartz, U. (2016). *Qualitative Inhaltsanalyse: Methoden, Praxis, Computerunterstützung (3., überarbeitete Auflage). Grundlagentexte Methoden*. Weinheim, Basel: Beltz Juventa. Retrieved from http://www.content-select.com/index.php?id=bib_view&ean=9783779943860

Leijten, M., & van Waes, L. (2013). Keystroke logging in writing research. *Written Communication, 30(3)*, 358–392. https://doi.org/10.1177/0741088313491692

Leki, I., & Carson, J. (1997). "Completely different worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31(1)*, 39. https://doi.org/10.2307/3587974

Li, J. (2014). The role of reading and writing in summarization as an integrated task. *Language Testing in Asia, 4(1)*, 3. https://doi.org/10.1186/2229-0443-4-3

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Linacre, J. M. (2021). *A user's guide to FACETS: Rasch-model computer programs*. Chicago, IL. Retrieved from Winsteps.com

Ling, G. (2017a). Are TOEFL iBT® writing test scores related to keyboard type? A survey of keyboard-related practices at testing centers. *Assessing Writing, 31*, 1–12. https://doi.org/10.1016/j.asw.2016.04.001

Ling, G. (2017b). Is writing performance related to keyboard type? An investigation from examinees' perspectives on the TOEFL iBT. *Language Assessment Quarterly, 14(1)*, 36–53. https://doi.org/10.1080/15434303.2016.1262376

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for rater training. *Language Testing, 12*, 54–71. https://doi.org/10.1177%2F026553229501200104

Marks, D. (2015). Prüfen sprachlicher Kompetenzen internationaler Studienanfänger an deutschen Hochschulen: Was leistet der TestDaF? *Zeitschrift für Interkulturellen Fremdsprachenunterricht, 20(1)*, 21–39.

McNamara, T. (1996). *Measuring second language performance*. London, New York: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp.13-103). New York, NY: American Council on Education.

Michel, M., Révész, A., Lu, X., Kourtali, N.-E., Lee, M., & Borges, L. (2020). Investigating L2 writing processes across independent and integrated tasks: A mixed-methods study. *Second Language Research, 36(3)*, 307–334. https://doi.org/10.1177/0267658320915501

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centered design*. ETS Research Report Series.

Moeller, A. J., Creswell, J. W., & Saville, N. (Eds.). (2016). *Second language assessment and mixed methods research*. Cambridge: Cambridge University Press.

Murray, D. (1982). *Learning by teaching*. Montclair, New Jersey: Boynton/Cook.

Norris, J., & Drackert, A. (2018). Test review: TestDaF. *Language Testing, 35(1)*, 149–157. https://doi.org/10.1177/0265532217715848

O'Connor, C., & Joffe, H. (2020). Intercoder Reliability in qualitative research: Debates and practical guidelines. *International Journal of Qualitative Methods*, 19, 160940691989922. https://doi.org/10.1177/1609406919899220

Ohkubo, N. (2009). Validating the integrated writing task of the TOEFL internet-based test (iBT): Linguistic analysis of test takers' use of input material. *Melbourne Papers in Language Testing, 14(1)*, 1–31.

Pecorari, D., & Petrić, B. (2014). Plagiarism in second-language writing. *Language Teaching, 47(3)*, 269–302. https://doi.org/10.1017/S0261444814000056

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13(2)*, 111–129. https://doi.org/10.1016/j.asw.2008.07.001

Plakans, L. (2009a). Discourse synthesis in integrated second language writing assessment. *Language Testing, 26(4)*, 561–587. https://doi.org/10.1177/0265532209340192

Plakans, L. (2009b). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes, 8(4)*, 252–266. https://doi.org/10.1016/j.jeap.2009.05.001

Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44(1)*, 185–194. https://doi.org/10.5054/tq.2010.215251

Plakans, L. (2013). Assessment of integrated skills. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–8). Malden, Mass., Hoboken, USA: Wiley-Blackwell; John Wiley.

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17(1)*, 18–34. https://doi.org/10.1016/j.asw.2011.09.002

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

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36(2)*, 161–179. https://doi.org/10.1177/0265532216669537

Plakans, L., Liao, J.-T., & Wang, F. (2019). "I should summarize this whole paragraph": Shared processes of reading and writing in iterative integrated assessment tasks. *Assessing Writing, 40*, 14–26. https://doi.org/10.1016/j.asw.2019.03.003

Révész, A., Michel, M., & Lee, M. (2017). *Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory* (IELTS Research Reports No. 2017/3).

Saldaña, J. (2016). *The Coding manual for qualitative researchers* (3rd ed.). Los Angeles: SAGE.

Sanders, T. J. M., & Schilperoord, J. (2006). Text structure as a window on the cognition of writing: How text analysis provides insights in writing products and writing processes. In C. M. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 386–402). New York, London: The Guilford Press.

Sawaki, Y., Quinlan, T., & Lee, Y-W. (2013). Understanding Learner Strengths and Weaknesses: Assessing Performance on an Integrated Writing Task. *Language Assessment Quarterly, 10(1)*, 73-95. https://doi.org/10.1080/15434303.2011.633305

Schnotz, W., Wagner, I., Zhao, F., Ullrich, M., Horz, H., McElvany, N., Ohle, A., & Baumert, J. (2017). Development of dynamic usage of strategies for integrating text and picture information in secondary schools. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Methodology of educational measurement and assessment. Competence assessment in education: Research, models and instruments* (pp. 303–313). Springer International Publishing.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge University Press.

Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21(2)*, 171–200. https://doi.org/10.1177/0741088303262846

Shin, S.-Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing, 32(2)*, 259–281. https://doi.org/10.1177/0265532214560257

Spivey, N. N. (1990). Transforming texts: Constructive processes in reading and writing. *Written Communication, 7(2)*, 256–287.

Spivey, N. N., & King, J. R. (1989). Readers as writers composing from sources. *Reading Research Quarterly, XXIV(1)*, 7–26.

Stezano Cotelo, K. (2003). Die studentische Seminararbeit: Studentische Wissensverarbeitung zwischen Alltagswissen und wissenschaftlichem Wissen. In K. Ehlich & A. Steets (Eds.), *Wissenschaftlich schreiben: Lehren und lernen* (pp. 87–114). Berlin, New York: W. De Gruyter.

Suvorov, R., & Hegelheimer, V. (2014): Computer-assisted language testing. In A. J. Kunnan (Ed.). The Companion to Language Assessment (pp. 593–613). Chichester, UK: Wiley-Blackwell.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge: Cambridge University Press.

van Dijk, T. A. (1979). Relevance assignment in discourse comprehension. *Discoures Processes, 2*, 113–126.

Wang, P. (2018). *Investigating test-takers' cognitive processes while completing an integrated reading-to-write task: Evidence from eye-tracking, stimulated recall and questionnaire*. Unpublished Doctoral Thesis. Lancaster University, UK.

Wang, J., Engelhard, G., Raczynski, K., Song, T., & Wolfe, E. (2017). Evaluating rater accuracy and perception for integrated writing assessments using a mixed-methods approach. *Assessing Writing, 33*, 36–47.

Watanabe, Y. (2001). *Read-to-write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions.* Unpublished doctoral dissertation. University of Hawaii.

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

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9(1)*, 27–55.

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing*, 21(2), 118–133. https://doi.org/10.1016/j.jslw.2012.03.004

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach. Research and practice in applied linguistics*. Basingstoke: Palgrave Macmillan.

Weir, C., O'Sullivan, B., Yan, J., & Bax, S. (2007). *Does the computer make a difference? The reaction of candidates to a computer-based versus a traditional hand-written form of the IELTS Writing component: effects and impact* (IELTS Research Reports No. 7). Canberra, Manchester, U.K.

Wette, R. (2017). Source text use by undergraduate post-novice L2 writers in disciplinary assignments: Progress and ongoing challenges. *Journal of Second Language Writing, 37*, 46–58. https://doi.org/10.1016/j.jslw.2017.05.015

Wolfersberger, M. (2013). Refining the construct of classroom-based writing-from-readings assessment: The role of task representation. *Language Assessment Quarterly, 10(1)*, 49–72. https://doi.org/10.1080/15434303.2012.750661

Xi, X. (2005). Do visual chunks and planning impact the overall quality of oral descriptions of graphs? *Language Testing, 22(4)*, 463–508. https://doi.org/10.1191%2F0265532205lt305oa

Xi, X. (2010). Aspects of performance on line graph description tasks: Influenced by graph familiarity and different task features. *Language Testing, 27(1)*, 73–100. https://doi.org/10.1177%2F0265532209346454

Yang, H.-C. (2014). Toward a model of strategies and summary writing performance. *Language Assessment Quarterly, 11(4)*, 403–431. https://doi.org/10.1080/15434303.2014.957381

Yang, H.-C. (2016). Describing and interpreting graphs: The relationships between undergraduate writer characteristics and academic graph writing performance. *Assessing Writing, 28(1)*, 28–42. https://doi.org/10.1016/j.asw.2016.02.002

Yang, H.-C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly, 46(1)*, 80–103. https://doi.org/10.1002/tesq.6

Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing, 14(2)*, 116–137.

Yu, G. (2013a). From integrative to integrated language assessment: Are we there yet? *Language Assessment Quarterly, 10(1)*, 110–114. https://doi.org/10.1080/15434303.2013.766744

Yu, G. (2013b). The use of summarization tasks: Some lexical and conceptual analyses. *Language Assessment Quarterly, 10(1)*, 96–109. https://doi.org/10.1080/15434303.2012.750659

Yu, G., He, L., & Isaacs, T. (2017). *The cognitive processes of taking IELTS Academic Writing Task 1: An eye-tracking study* (IELTS Research Reports).

Yu, G., Rea-Dickins, P., & Kiely, R. (2007). *The cognitive processes of taking IELTS Academic Writing Task 1* (IELTS Research Reports No. 11).

Zhao, F., Schnotz, W., Wagner, I., & Gaschler, R. (2014). Eye tracking indicators of reading approaches in text-picture comprehension. *Frontline Learning Research, 6*, 46–66. https://doi.org/10.14786/flr.v2i4.98

Zimmermann, S. (2009). Falsche Vorbereitung? Erkenntnisse aus Teilnehmerleistungen der Prüfungsteile Mündlicher und Schriftlicher Ausdruck im Test Deutsch als Fremdsprache (TestDaF). In X. Yu (Ed.), *TestDaF-Training und Studienvorbereitung: Beiträge zur chinesisch-deutschen Fachkonferenz: „TestDaF-Training und Studienvorbereitung" vom 11. bis 12. Oktober 2008 am Deutschkolleg der Tongji-Universität Shanghai* (pp. 63–83). München: Iudicium.

Zimmermann, S. (2020a). "Das ist doch Leseverstehen!": Eine empirische Untersuchung zum Konstrukt von integrierten Schreibaufgaben. In A. Drackert, M. Mainzer-Murrenhoff, A. Soltyska, & A. Timukova (Eds.), *Testen bildungssprachlicher Kompetenzen und akademischer Sprachkompetenzen.: Zugänge für Schule und Hochschule* (pp.187-123). Frankfurt/M.: Peter Lang.

Zimmermann, S. (2020b). "Das ist Zeitverlust für mich, den Text wieder lesen": Einblicke in das Schreiben von Zusammenfassung in der Fremdsprache Deutsch. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 25(2), 111–133.

APPENDIX

Appendix A: Consent form

Einwilligungserklärung zur Datenerhebung

Ich habe die Datenschutzerklärung zur Kenntnis genommen und stimme der Verarbeitung und Speicherung meiner personenbezogenen Daten im Rahmen der Erprobung neuer TestDaF-Aufgaben und des Eye-Tracking zu. Ich kann die Einwilligung jederzeit widerrufen. Die Rechtmäßigkeit der Verarbeitung meiner Daten bis zum Widerruf bleibt hiervon unberührt.

Name _____

Vorname _____

Unterschrift _____

Einwilligungserklärung zu Videoaufnahme

Im Rahmen des Eye-Tracking eingesetzte Videoaufnahme dient ausschließlich dem Zweck, Ihre Kopfbewegungen aufzuzeichnen und so Hinweise auf Ihr Blickverhalten außerhalb des Computerbildschirms zu bekommen.

Die Videoaufzeichnung wird sicher aufbewahrt und vor unbefugten Zugriffen geschützt. Eine Weitergabe der Videoaufzeichnung an Dritte erfolgt nicht. Die Videoaufnahme wird umgehend gelöscht, sofern der Zweck der Aufnahme erfüllt ist.

Ich bin damit einverstanden, dass die Gesellschaft für Akademische Studienvorbereitung und Testentwicklung e.V. (g.a.s.t.) bzw. das TestDaF-Institut die Videoaufnahme während des Eye-Tracking anfertigt. Ich kann die Einwilligung jederzeit widerrufen.

Name _____

Vorname _____

Unterschrift _____

Appendix B: Guideline stimulated recall

VOR DEM EYE-TRACKING

1. Begrüßung

   Schön, dass Sie da sind und sich bereit erklärt haben, an meiner Studie mitzumachen.
   2-3 Sätze Smalltalk, z.B. Haben Sie alles gut gefunden?

2. Ziel der Studie

   In dieser Studie interessiert mich, wie Sie bei der Bearbeitung einer Schreibaufgabe am Computer vorgehen. Um das herauszufinden, werde ich Ihre Blickbewegungen auf dem Bildschirm mithilfe einer speziellen Software aufzeichnen. Diese läuft im Hintergrund, so dass Sie davon nichts mitbekommen und sich ganz auf die Aufgabe konzentrieren können. Danach werde ich noch ein kurzes Interview mit Ihnen führen.

3. Erläuterung des Ablaufs

   Wir werden nun zunächst eine sogenannte Kalibrierung durchführen. Die ist notwendig, damit die Software Ihre Blickbewegungen am Bildschirm so genau wie möglich aufzeichnen kann. Wenn diese erfolgreich war, beginnen wir mit der Schreibaufgabe. Ihre Blickbewegungen am Bildschirm werden dabei aufgezeichnet. Zusätzlich zeichnet diese Videokamera auf, wohin Sie schauen, wenn Sie nicht auf den Bildschirm sehen.

   Nach Bearbeitung der Aufgaben werden wir uns gemeinsam ein paar Ausschnitte aus der Aufzeichnung ansehen und ich werde Sie bitten, sich daran zu erinnern, was Sie zu diesem Zeitpunkt gedacht haben. Ich werde dieses Gespräch ebenfalls aufzeichnen.

   Ganz zum Schluss möchte noch Ihre Tippgeschwindigkeit am Computer testen. Dafür möchte ich Sie bitten, einen kurzen Text am Bildschirm abzutippen. Davon wird es keine Bildschirm- oder Videoaufzeichnung geben. Ich benötige die Ergebnisse dieses kurzen Tests aber, um einzuschätzen, wie sicher und schnell Sie mit dem Schreiben am Computer zurechtkommen.

   Soweit alles klar? Haben Sie noch Fragen?

4. Datenschutz und Einverständniserklärung

   Ich werde im Rahmen dieser Studie personenbezogene Daten von Ihnen erheben, z.B. Ihren Namen, Ihr Herkunftsland, Ihr Alter, Ihr Muttersprache, Ihre Ergebnisse, Ihre Eye-Tracking-Daten, Tonaufnahmen und auch eine Videoaufzeichnung. Bitte sagen Sie mir, ob Sie mit der Videoaufzeichnung einverstanden sind. Wenn nicht, schalte ich die Kamera aus.

Ich möchte Sie nun bitten, mir eine Einverständniserklärung zu unterzeichnen. Damit erklären Sie sich einverstanden, dass ich Ihre Daten in anonymisierter Form für Forschungszwecke verwenden darf.

5. Kalibrierung

   Beginnen wir nun mit der Kalibrierung.

   Sitzen Sie bequem? Sie sehen gleich einen Punkt in der Mitte des Bildschirms. Bitte fixieren Sie diesen mit den Augen. Wenn der Punkt sich anfängt zu bewegen, folgen Sie ihm mit Ihren Augen über den Bildschirm. Bewegen Sie möglichst nicht den Kopf oder Ihren Körper.

   Sind Sie bereit? Dann geht es jetzt los.

6. Schreibaufgabe

   Wir beginnen jetzt mit der Schreibaufgabe.

   Bitte denken Sie daran, sich in den nächsten 30 Minuten möglichst wenig zu bewegen, damit Ihre Blickbewegungen möglichst genau aufgezeichnet werden.

   Klicken Sie auf „Starten“, wenn Sie bereit sind, mit der Bearbeitung der Aufgabe zu beginnen.

NACH DEM EYE-TRACKING

7. Allgemeine Schwierigkeit und Textverständnis

   a) Wie war das für Sie?

   b) Was war besonders schwierig? Was war leicht?

   c) Können Sie mir kurz zusammenfassen, worum es in dem Text und der Grafik ging?

      - je nach Schreibaufgabe: konkrete Fragestellung wiederholen und nach Informationen dazu fragen

8. Erläuterung Stimulated Recall

   Wir schauen uns nun Ausschnitte aus einem Video Ihrer Schreibsitzung an.

   Mich interessiert, was Sie während des Schreibens gedacht haben. Wir können in dem Video sehen, wohin Sie auf dem Bildschirm schauen, was Sie schreiben, und auch wohin Sie schauen, wenn Sie nicht auf den Monitor sehen. Aber wir wissen nicht, warum Sie gerade an eine bestimmte Stelle geschaut haben, und was Sie dabei gedacht haben.

   Ich würde Sie daher bitten, sich einige Ausschnitte mit mir gemeinsam anzusehen und mir zu sagen, was zu dem Zeitpunkt Ihre Gedanken waren.

Wenn Sie sich das Video ansehen, versuchen Sie sich an den Prozess des Schreibens zurückzuerinnern. Mich interessiert das, was Sie während des Schreibens gedacht haben, und nicht das, was Sie gerade jetzt denken. Es ist wichtig, dass Sie mir alle Ihre Gedanken mitteilen, egal wie abwegig oder albern Sie Ihnen vorkommen.

Ich werde dieses Gespräch ebenfalls aufzeichnen.

Alles klar soweit? Haben Sie noch Fragen?

9. Videoausschnitte

Jetzt gehen wir mal in das Video.

a) Bestimmte Stellen allgemein betrachten, z.B. unterschiedliche Phasen des Schreibprozesses wie das Lesen der Aufgabenstellung, Lesen des Inputtextes

- Können Sie mir sagen, was Sie an dieser Stelle gedacht haben?
- Warum haben Sie hier x/y gemacht?
- Gibt es noch etwas, das Ihnen einfällt?

b) Bestimmte Stellen problematisieren, z.B. unterschiedliche längere Pausen im Schreibprozess, wiederholtes Lesen des Inputtextes oder des eigenen bisher geschriebenen Textes, Blicke zur Uhr oder zur Wörterzählung

- Ich sehe, dass Sie hier mit dem Schreiben aufgehört haben und eine längere Pause gemacht haben. Was haben Sie zu dem Zeitpunkt gedacht?
- Ich sehe, dass Sie hier Änderungen (Einfügen, Entfernen, Umstellen) an Ihrem Text vorgenommen haben. Können Sie mir sagen, was Sie zu dem Zeitpunkt gedacht haben?
- Sie haben hier wiederholt den Text und die Grafik angeschaut. Können Sie mir sagen, was Sie zu dem Zeitpunkt gedacht haben?
- Ich sehe, dass Sie hier häufig auf die Uhr bzw. die Wörterzählung geschaut haben. Was haben Sie hier gedacht?
- Gibt es noch etwas, das Ihnen einfällt?

c) Ergänzungen von Seiten der Teilnehmer
Haben Sie noch weitere Anmerkungen?

Danke. Ich werde jetzt die Aufnahme beenden.

10. Type Speeding Test

Zum Schluss würde ich Sie noch bitten, für two Minuten einen kurzen Text am Bildschirm abzutippen. Dies hilft bei der Einschätzung, wie sicher und schnell Sie beim Schreiben am Computer sind.

Die Zeit läuft, sobald Sie anfangen zu schreiben.

Nach Ende des Tests erscheint ein Feedback. Bitte klicken Sie die Seite nicht weg oder schließen den Browser.

11. Ende und Formalia

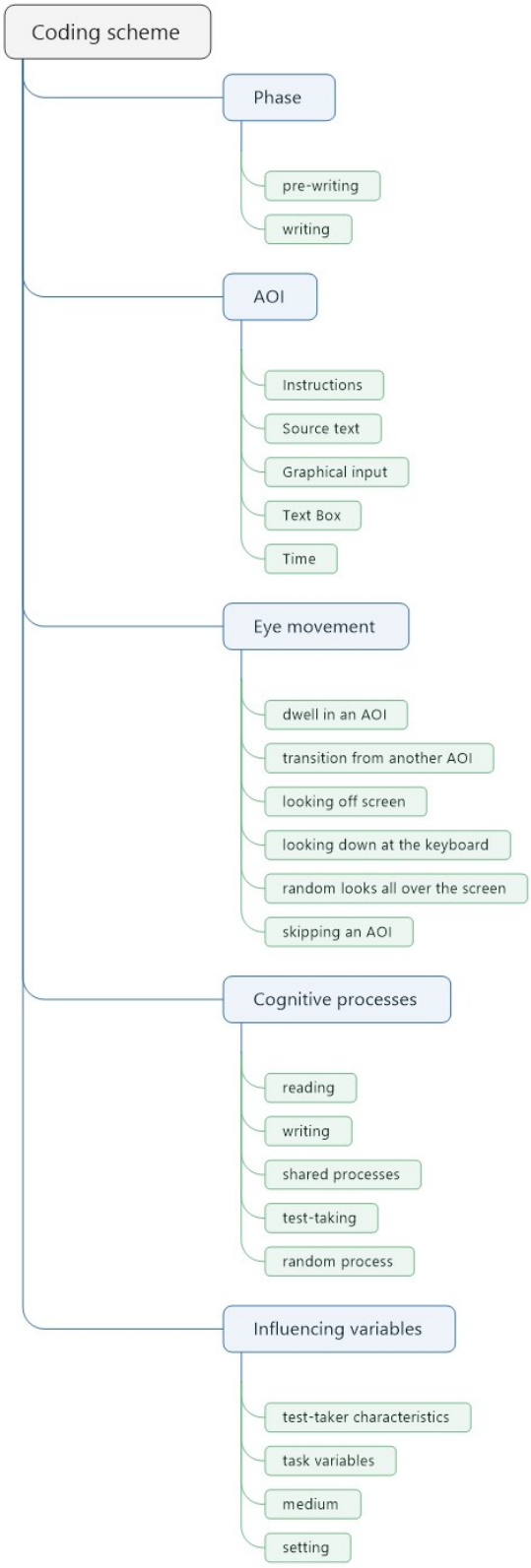
Vielen Dank, dass Sie so lange durchgehalten haben.

Ich würde Sie bitten, mir kurz auf dieser Liste zu bestätigen, dass Sie die vereinbarte Aufwandsentschädigung erhalten haben.

Appendix C: Validation results overview

Participant	Right Eye Deviation X [°]	Right Eye Deviation Y [°]	Left Eye Deviation X [°]	Left Eye Deviation Y [°]
1-02	0.36	0.49	0.43	0.32
1-03	0.20	0.33	0.25	0.17
1-04	0.41	0.35	0.49	0.44
1-05	0.25	0.50	0.45	0.78
1-06	0.04	0.21	0.22	0.21
1-07	0.43	0.46	0.17	0.24
1-09	0.30	0.27	0.34	0.30
1-10	0.31	0.47	0.48	0.45
2-01	0.49	0.28	0.25	0.67
2-03	0.57	0.59	0.40	0.54
2-04	0.22	0.23	0.15	0.16
2-05	0.31	0.20	0.23	0.33
2-06	0.54	1.60	0.61	0.47
2-07	0.29	0.61	0.14	0.26
2-08	0.60	0.51	0.35	0.25
2-09	0.39	0.51	0.44	0.71
2-10	0.46	0.85	0.48	0.26

Appendix D: Coding scheme stimulated recalls



Appendix E: Inter-coder agreement

Code	Agreement (%)	Cohen's Kappa
Phase	**93.10**	**.85**
Pre-writing	93.46	.81
Writing	92.75	.85
AOI	**90.91**	**.59**
Graphical input	94.36	.57
Instructions	91.46	.58
Source text	92.43	.74
Text box	80.06	.30
Time	96.22	.79
Eye movement	**87.62**	**.43**
Dwell	74.67	.32
Looking off screen	100.00	1.00
Random looks	96.07	.00
Skipping an AOI	92.68	.36
Transition	74.69	.42
Cognitive processes	**96.19**	**.47**
Random process	100.00	1.00
Reading	95.92	.37
Shared processes	96.56	.55
Test-taking	93.77	.55
Writing	96.55	.36
Influencing variables	**92.40**	**.37**
Medium	99.32	.75
Setting	100.00	1.00
Task variables	74.45	.28
Test taker characteristics	93.08	.34

Appendix F: Comparison of viewing behavior across sets

	Set 1	Set 2	Mann-Whitney-U-Test		
	Mdn	***Mdn***	***U (Z)***	***p***	***r***
pre-writing time	4:19	4:00	33.000 (-.289)	.815	
writing time	25:45	25:30	32.000 (-.385)	.743	
Dwell time in different AOIs (% of total task duration)					
Instructions	3.75	2.90	30.000 (-.578)	.606	
Source Text	22.35	23.00	40.500 (.433)	.673	
Graphical Input	8.50	5.50	16.000 (-1.925)	.059	
Text Box	35.85	24.10	25.000 (-1.058)	.321	
Time	.30	.30	29.500 (-.636)	.541	
Revisits in different AOIS (% of all revisits)					
Instructions	4.75	4.76	34.500 (-.144)		
Source Text	20.81	25.66	44.000 (.770)		
Graphical Input	21.80	17.95	18.000 (-1.732)		
Text Box	44.81	43.39	23.000 (-1.251)		
Time	4.74	5.82	42.000 (.577)		
Transitions between the different AOIs (% of all transitions)					
Instructions – Source Text	.71	.91	50.000 (1.350)	.200	
Instructions – Graphical Input	.63	.88	36.000 (.000)	1.000	
Instructions – Text Box	1.99	2.19	36.000 (.000)	1.000	
Instructions – Time	.00	.00	33.500 (-.298)	.815	
Source Text – Instructions	.77	.88	42.500 (.626)	.541	
Source Text – Graphical Input	4.07	3.39	41.000 (.481)	.673	
Source Text – Text Box	11.99	10.79	44.000 (.770)	.481	
Source Text – Time	2.05	2.24	44.000 (.770)	.481	

Graphical Input – Instructions	1.12	.67	29.000 (-.674)	.541	
Graphical Input – Source Text	4.47	3.39	30.000 (-.577)	.606	
Graphical Input – Text Box	20.69	18.75	24.000 (-1.155)	.277	
Graphical Input – Time	.29	.00	26.500 (-.965)	.370	
Text Box – Instructions	1.04	.88	39.000 (.292)	.815	
Text Box – Source Text	10.81	13.00	41.000 (.481)	.673	
Text Box – Graphical Input	22.46	17.45	22.000 (-1.347)	.200	
Text Box – Time	1.33	1.69	31.000 (-.481)	.673	
Time – Instructions	.28	.00	26.500 (-1.023)	.370	
Time – Source Text	1.48	3.07	64.000 (2.694)	.006	.65
Time – Graphical Input	.37	.85	36.000 (.000)	1.000	
Time – Text Box	3.20	3.59	34.000 (-.192)	.888	

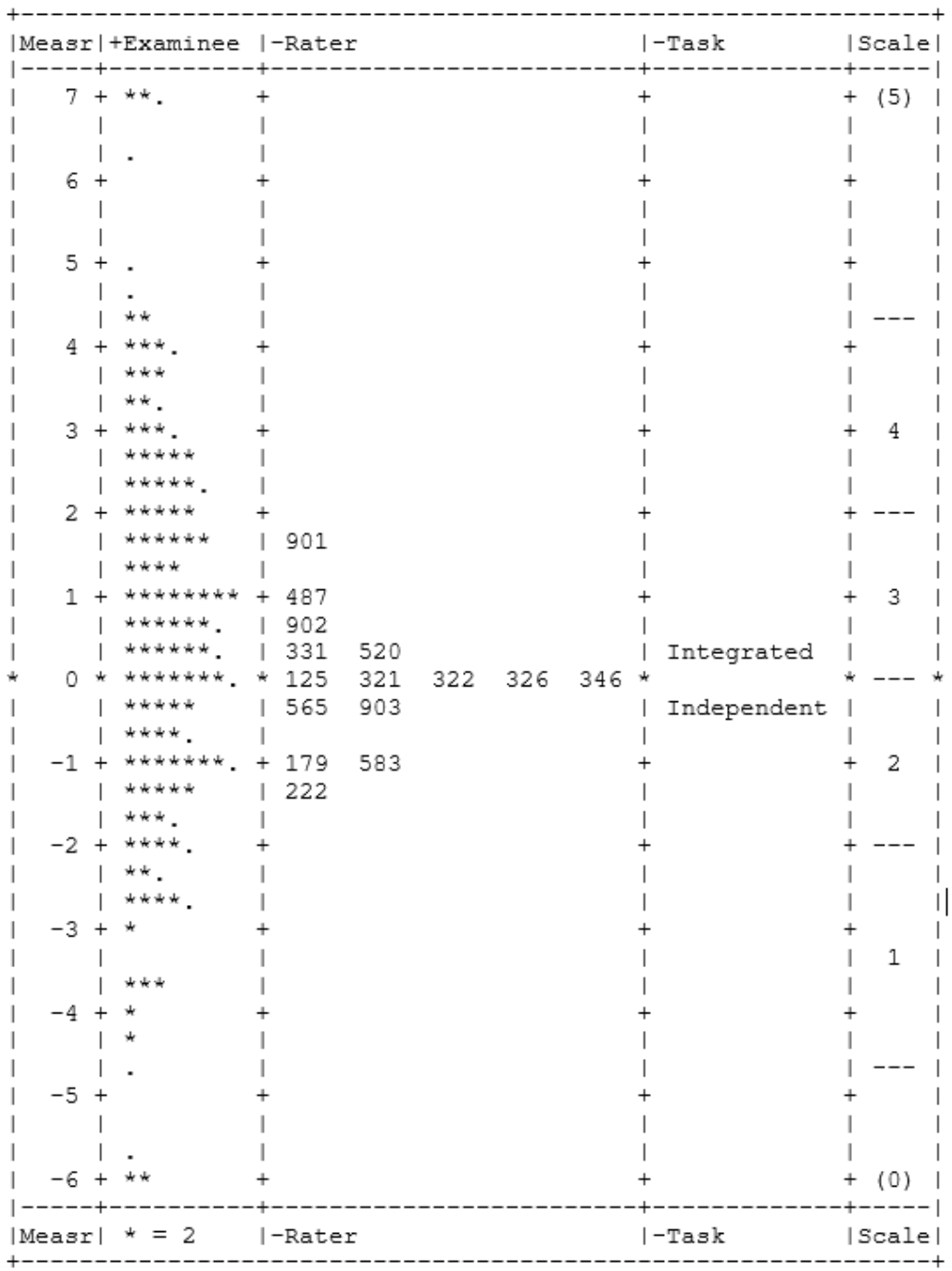
Note. Set 1: N=8; Set 2: N=9.

Appendix G: Integrated writing scores for embedded sample

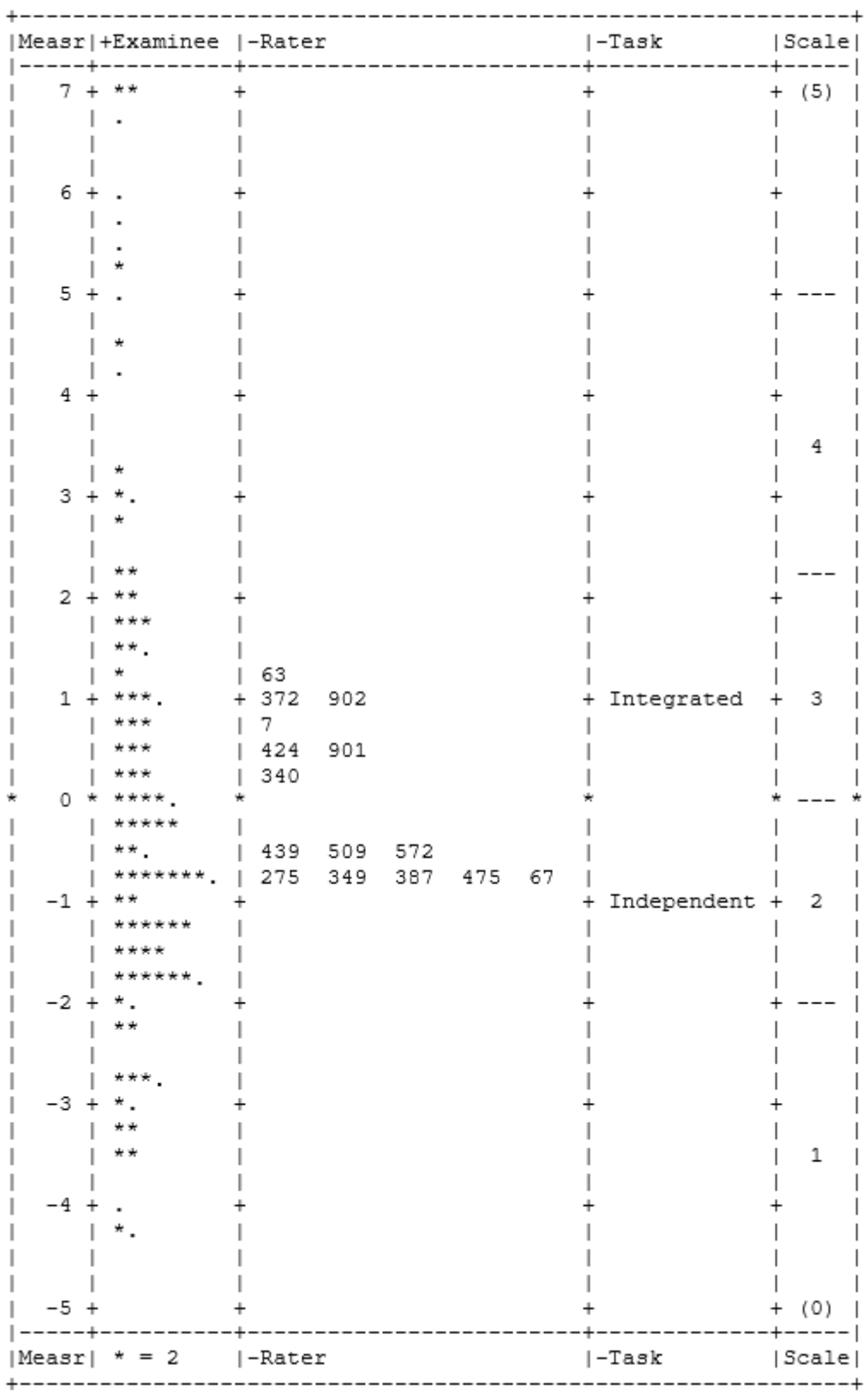
Participant	Observed average	Fair average
1-02	3.50	3.60
1-03	2.00	1.68
1-04	3.50	3.08
1-05	1.00	1.03
1-06	2.00	1.71
1-07	4.50	4.46
1-09	2.50	2.25
1-10	2.50	2.33
2-01	2.50	2.43
2-02	1.00	1.04
2-03	2.00	1.86
2-04	4.00	3.74
2-05	2.00	2.50
2-06	2.00	1.56
2-07	2.50	2.43
2-08	1.50	1.91
2-09	2.50	2.24
2-10	2.00	1.86
2-11	1.00	1.18

Appendix H: Wright maps from the many-facet rating scale analysis across tasks in Set 1 and Set 2

Set 1



Set 2


|Measr|+Examinee |-Rater |-Task |Scale|
7 (5)
6
5 ---
4
4
3

2
1 63 372 902 7 424 901 340 Integrated 3
0 ---
-1 439 509 572 275 349 387 475 67 Independent 2
-2 ---
-3
1
-4
-5 (0)
|Measr| * = 2 |-Rater |-Task |Scale|

Appendix I: Comparison low- vs. high-scoring participants

	low (N=13)	**high** (N=6)	**Mann-Whitney U Test**		
	Mdn	***Mdn***	***U (Z)***	***p***	***r***
Origin of information					
Source Text			50.500 (1.012)	.323	
Graphical Input			39.000 (.000)	1.000	
unknown			22.000 (-1.519)	.152	
Relevance and accuracy of information					
relevant and correct			60.000 (1.874)	.072	
relevant, but minor issues			37.000 (-.176)	.898	
not relevant			36.500 (-.221)	.831	
false			33.000 (-.541)	.639	
major issues			15.000 (-2.229)	.036	.51
Paraphrase type					
exact copy			38.000 (-.123)	.966	
near copy			32.000 (-.650)	.579	
minimal revision			45.500 (.575)	.579	
moderate revision			28.000 (-.970)	.368	
substantial revision			48.500 (.903)	.416	