Using Task Features to Predict Item Difficulty and Item Discrimination in 3F Dutch Reading Comprehension Exams

Researcher: Nancy de Groot BSc

Supervisor: Erik Roelofs PhD

External (first) supervisor: Dr. Karen Keune

External (second) supervisor: Dr. Remco Feskens

Keywords: task features, predicting item difficulty, predicting item discrimination, reading comprehension

Word count: 19349

19-06-2023

Acknowledgements

Before diving into the contents of this study, I would like to express my gratitude for the support I have gotten during preparing, conducting, and completing this study. In particular, I would like to thank dr. Eric Roelofs, who inspired me by his contagious enthusiasm for assessment, and more specifically, assessment of reading comprehension. Moreover, I would like to thank Cito, the Dutch testing institute, for having me around as an intern conducting my master thesis and giving me a look behind the scenes. I would like to thank dr. Remco Feskens guided me through the use of R, and his support during the data analysis. Finally, I would like to thank dr. Karen Keune, who took on the role of daily supervisor at Cito.

Abstract

Reading comprehension is an important target skill in education. However, there has been criticism on the assessment of reading comprehension, mainly on the construct validity. This study attempts to consider some of the criticism, aiming to identify task features that either support or impede construct relevance, reflected by the psychometric quality indicators item difficulty and item discrimination. The main benefit in finding these task features is the potential increase in construct validity, and the connected increased coverage of the target skill, providing more valid assessment. This study focused on finding these predictive task features in 3F Dutch language comprehension exam items (N = 182), aimed at the highest reference level for vocational education, created by Cito and administered between 2015 to 2022. The focus was on finding (groups of) task features with predictive value for the item parameters from previous research, followed by analyses on the relation between task features and item parameters in the present item sample. Subsequently, the predictive value of task features for item parameters is analysed and discussed. The focus of the present study was on the lexical richness of the text, represented by the proportional complexity, based on concept maps. Key findings include text and text content as predictors for item parameters, and the degree to which the key is the best possible response (increased discriminative ability, and easier item). Regarding item difficulty, the number of plausible response options increased difficulty. Other predictors were found in the group of intrinsic task features and access skills.

Keywords: reading comprehension, predicting item difficulty, predicting item discrimination, predictive task features

Table of Contents

Acknowledgements 2
Abstract 3
Introduction 6
Theory-Driven Approach: Evidence-Centred Design Framework 7
Using Task Features as Item Parameter Predictors 8
The Present Context: 3F Dutch Reading Comprehension Exams 8
Theoretical Framework 10
Challenges In Reading Comprehension Assessment 10
Confronting challenges in reading comprehension assessment 10
Evidence-Centred Design Framework 11
Task Features Predicting Item Difficulty 12
2F Dutch comprehension exams 12
3F Dutch comprehension exams 13
Overview of (groups of) task features predicting item difficulty 14
Task Features Predicting Item Discrimination 15
2F Dutch comprehension exams 15
3F Dutch comprehension exams 16
Overview of (groups of) task features predicting item discrimination 16
The Use of Concept Maps to Represent the Situation Model 17
Methods 19
Selected Items and Associated Exam Data 19
Coding task features 21
Coding propositions deemed necessary to arrive at a correct response to the item 24
Coding scheme for task features 28
Results 34
Descriptive Statistics: Item Parameters and Task Features 34
Prevalence of proposition features 34
Prevalence of types of relationships 36
Correlation Analyses and Effect Sizes: Item Parameters and Task Features 39
Text Predicting Item Parameters 41
Prediction of Item Difficulty Using All Task Features 42

Prediction of Item Discrimination 45
Conclusion and Discussion 49
Predicting Item Difficulty 49
Predicting Item Discrimination 51
Text (Content) as Predictor for Item Parameters 52
Limitations of The Present Research and Suggestions for Further Research 53
Item sample 53
Methods 53
Analyses 54
General suggestions 55
References 56
Appendices 61
Annex 1: Description of the Specified Guidelines for the Reference Level 3F per Skill (College voor Toetsen en Examens, 2017) 61
Annex 2: Coding Scheme to Determine Task Features per Item in Microsoft Access, Created by Roelofs, Postulart, et al. (2021) 62
Annex 3: Results of Correlational Analyses and Analyses of Variance Between All Task Features and Item Parameters (N = 182) 63
Annex 4: Description of the Results of One-way Analyses of Variance (ANOVA) 65

Introduction

Reading Comprehension: Target Skill and its Assessment

The importance of reading comprehension as a target skill is emphasised by several researchers for different reasons (e.g., Alderson, 2000; Kendeou et al., 2014; Kendeou et al., 2016; Niklas et al., 2016; Paris & Hamilton, 2014; Snow, 2002). Snow (2002) found that learners who do not master reading comprehension, are prone to have difficulty throughout their work and education. Furthermore, the significance of reading comprehension for both academic success and long-term success in life is emphasised by Kendeou et al. (2016). Moreover, a basic level of reading comprehension is one of the necessities for developing 21st-century skills (Graesser, 2015). Goldman and Pellegrino (2015) state contemporary citizens should possess these skills to apply their knowledge to solve (new) 21st-century problems.

Particularly in the context high stakes assessment and for purposes of international comparisons, usually, the assessment of students' reading comprehension involves some form of a standardised reading comprehension ability test (Ozuru et al., 2008). These standardised reading comprehension ability tests frequently include several short text passages and different multiple-choice items regarding the contents of these text passages for the test takers to be answered. As performing well on these test entails that the test takers both read the text passages and answer items regarding the contents correctly, it is assumed these standardised tests are able to measure test takers' reading ability.

Lately, there has been criticism of the assessment of reading comprehension, where the focus is on the assumed decrease in construct validity. Construct validity is acknowledged broadly as a key quality of assessment in the field of measurement (Anderson et al., 1991; Snow, 2003) and represents the coverage of the target skill, as assessment should measure the stated instructional objectives (Downing, 2003).

Researchers posit that the measurement of the target skill reading comprehension has been reduced to relatively restricted types of items, which include short passage reading and responding to multiple-choice items, while searching and destroying false options (Rupp et al., 2006; Snow, 2003). When the latter is the case, forming a mental representation of the text and considering the intention of the author is not required. As a result, the construct validity of reading comprehension is threatened. Following this line of criticism, inferences about the test takers' levels of reading comprehension may no longer hold, or only reflect a reduced conception of reading comprehension, because of the limited task model behind reading comprehension items.

Therefore, research on the quality of reading comprehension assessment is necessary in order to identify those task features of reading comprehension items that either support or impede claims about test takers' reading comprehension skill. More specifically, the current study aimed to explore and classify item features related to psychometric evidence these items elicit, indicated by item difficulty and item discrimination, in a standardised reading comprehension ability test. Estimates of item difficulty and item discrimination are yielded through item analysis and are used to determine the quality of items quality in a systematic manner (Saswati, 2021). Firstly, item difficulty in the classical theory (as expressed by p-value) refers to the proportion of test takers that correctly responded to a certain item (e.g., Bibler Zaidi et al., 2018). A high value indicates a relatively high number of test takers answered the item correctly: the corresponding item is therefore a relatively easy item. Secondly, item discrimination, as represented by a discrimination index, rir-value in classical test theory and the beta-parameter in item response theory, informs about the evidence strength of an item (e.g., Bibler Zaidi et al., 2018). The higher the discrimination parameter, the better an item discriminates between test takers, who perform high and low respectively on the test (as used by Cito). As a minimal value required of item-rest correlations for maximum performance or cognitive tests, some rules of thumb mentioned are .20, .30 or .40 (Zijlmans et al., 2018).

Theory-Driven Approach: Evidence-Centred Design Framework

Past research has focused on a variety of issues regarding the validity of assessment, and in particular standardised reading comprehension tests (Ozuru et al., 2008). The focus of research was both on the texts (e.g., length and structure), and the corresponding items (more specifically, the item format). More recently, research has also included research from other fields, such as the field of technology, and the field of cognitive psychology (Mislevy et al., 2003; Mislevy & Haertel, 2007). In their evidence-centred design (ECD) framework, multiple interconnected models have to be aligned in order to ensure high-quality assessment (Mislevy et al., 2003). A limitation in one of the models can have an impact on the assumptions made regarding test takers' performance, thereby threatening claims about the test takers' target skill, and hence threatening assessment validity. In that sense, this framework can help to obtain insights about the factors that influence the quality of assessment.

In the current study, the ECD framework was applied to investigate the assessment of reading comprehension. Within this framework, systematic construction of a cognitively based task model is stressed in order to cover the target skill. This framework has been applied by researchers in different contexts to evoke mental activities based on tasks with different purposes in varying contexts (Roelofs, Emons, et al., 2021). Applying the ECD framework will enable evidentiary arguments in order to underpin the assessment of reading comprehension as a target skill. More

specifically, in this study task features of reading comprehension items are identified and coded, as part of a coding system that represents and ad posteriori description of the underlying task model. Ad posteriori includes both task features that were intentionally used a manipulated (such as test types and response) and features that were implicitly used or features that slipped unintentionally (such as incomplete key responses). As stated before, the impact of these task features on item difficulty and item discrimination is studied.

Using Task Features as Item Parameter Predictors

Task features, or task model variables, have been found to affect item difficulty and item discrimination (Roelofs, Emons, et al., 2021). Task model variables are described as those task features that "are important for designing, calibrating, selecting, executing, and scoring it (the task)" (Mislevy et al., 2003, p. 23). The variables are related to describing the task itself, including the task environment. Identifying the task features and their positive or negative impact on the psychometric quality of an item, will help test constructors to enhance quality of items (Roelofs, Emons, et al., 2021). This implicates items that have a relatively high evidence strength, across various levels of difficulty. Enhanced quality items will in the end also be beneficial for test takers, as their reading comprehension skill is measured more validly and will provide bespoke information to act on. In the current study part of the criticism on the assessment of reading comprehension is addressed, by investigating the impact of those task features on item difficulty and item discrimination, which are assumed to have negative effects on construct validity.

The Present Context: 3F Dutch Reading Comprehension Exams

Following a theory informed approach, the present study focuses on 3F Dutch reading comprehension exams for Dutch vocational education in the Netherlands, for which test data have been collected for almost ten years. Dutch reading comprehension exams are constructed under supervision of the Dutch Committee for Testing and Evaluation (CvTE) by Cito, the central institute for assessment development in the Netherlands (Cito, 2022a). These central exams are administered online at five moments throughout the school year. Dutch language is a mandatory school subject at all four levels of Dutch vocational education (Cito, 2022b). The central exam for Dutch language for Dutch vocational education level 4 (the highest level) aims to measure mastery of the third level (3F) of language competence, as defined in the Dutch framework that encompasses four levels of language competence. The framework was set up by the Dutch government in order to guide and improve language education at Dutch schools (Rijksoverheid, 2022). The so-called reference level 3F represents a proficient language user, who is able to be a self-employed professional, or a starting student in higher professional education (College voor Toetsen en Examens, 2017). Moreover, the Dutch Committee for Testing and Evaluation has specified general learning outcomes for language

users on the reference level 3F on several target skills, including understanding, interpreting, evaluating, summarising, and looking up, which can be found in Annex 1.

Previous researchers developed a coding system for the items of 3F Dutch reading comprehension exams using different task features (Roelofs, Postulart, et al., 2021). This study will build on this work, by extending the coding system with additional task features, specifically those regarding the text information. In doing so, this study aims to explore the relations between task features on the one hand and the item difficulty and item discrimination on the other hand, in order to evaluate and subsequently optimize the quality of reading comprehension assessment. To reach this aim, the research question is: "*To what extent do (groups of) task features predict item difficulty and item discrimination in 3F Dutch language reading comprehension exams in secondary vocational education in the Netherlands?*"

Building up to formulating an answer to this main research question, first, the focus will be on *what (groups of) task features can be found to predict item difficulty and item discrimination in Dutch language reading comprehension exams.* After an extensive description of previous research in both 2F and 3F Dutch language comprehension exams, the focus will be on the relation between these task features and the item parameters, as analysed in the present study. The focus is to find out *to what extent the (groups of) task features are related to item difficulty and item discrimination in 3F Dutch language reading comprehension exams.* The relations found between both (groups of) task features and item parameters will help to give a clue as to which task features might be predictive for either item difficulty, item discrimination, or both. The study will then focus on answering the main research question: "*To what extent do (groups of) task features predict item difficulty and item discrimination in 3F Dutch language reading comprehension exams in secondary vocational education in the Netherlands?*"

Theoretical Framework

Challenges In Reading Comprehension Assessment

There are several challenges in developing satisfactory reading comprehension assessments (Snow, 2003). This section focuses on two main challenges in assessing reading comprehension, focused on the complexity of this target skill and related research challenges. A first challenge in reading comprehension assessment is that reading comprehension is a complex and multicomponent skill (e.g., Kendeou et al., 2016; Perfetti & Stafura, 2014). Reading comprehension comprises a complex combination of linguistic, cognitive, and metacognitive processes that are used by the reader to construct meaning from written texts (Van den Broek et al., 1995). To understand written text, a reader must draw inferences using relevant previous knowledge, recognise the text's structure and consider the objectives and motivations of a text's author (Graesser, 2015). Aiming to measure test takers' reading comprehension, Kintsch and Van Dijk (1978) describe how all these processes result in a mental representation reflecting the overall meaning of the text, often referred to as the 'situation model'. Snow (2003) mentions there is a need to identify reading comprehension processes, such as inferencing, and the integration of new information with present knowledge, and distinguish them from other skills involved with reading comprehension. These skills are for example one's knowledge of vocabulary, knowledge about the domain, and the decoding skills.

As a second challenge in reading comprehension assessment, it is mentioned that the target domain itself is complicated, which makes it difficult to achieve construct validity (Snow, 2003). Snow (2003) also mentions that it is unlikely that all researchers and practitioners will agree fully on the definition of "real comprehension", which causes choosing or creating comprehension measures to be very challenging. In addressing the reading comprehension skill, the present study adopted the following definition of reading comprehension: "the process of simultaneously extracting and constructing meaning through interaction and involvement with written language" (Snow, 2002, p. 14).

Confronting challenges in reading comprehension assessment

Messick (1996) discussed how to confront issues regarding the complicated and multifaceted character of validity in testing complex psychological constructs (e.g., reading comprehension). In particular, six dimensions of construct validity are highlighted: (1) content, (2) substantive, (3) structural, (4) generalisability, (5) external, and (6) consequential. Regarding these six dimensions, it is argued all dimensions are needed to improve understanding of strengths and weaknesses in assessment.

In line with these different dimensions of construct validity, researchers have examined issues related to validity, frequently aimed at the item format used: multiple-choice items. For various reasons, this item format is considered detrimental for making valid claims about the target skill reading comprehension. Several researchers claims that multiple-choice items focus only on a certain passage of the text (Rupp et al., 2006; Snow, 2003). These type of items fall short in both accurately eliciting the targeted comprehension processes as these occur in reading in practice, and in how these address all necessary skills and abilities for comprehension (Snow, 2003). Ozuru et al. (2008) mention two approaches regarding research into the validity of item formats (e.g., multiple-choice items, open items) in measuring reading comprehension: a statistical approach and a process oriented (a more experimental) approach. On the one hand, research that applies a statistical approach focused on explaining the variance in reading comprehension item difficulty, as a function of different item formats. On the other hand, research using the process-oriented approach apply think-aloud procedures. In these studies, test takers verbalise their thoughts while answering items in different formats, thereby highlighting mental processes that are involved in solving these tasks and evaluating to which these represent processes deemed essential for reading comprehension. The latter approaches applies knowledge from the field of cognitive psychology (Mislevy et al., 2003; Mislevy & Haertel, 2007).

Evidence-Centred Design Framework

In this study, the evidence-centred design (ECD) framework is used to study the 3F Dutch reading comprehension exams. This framework has been previously applied to Dutch driving theory exams (Roelofs, Emons, et al., 2021). The ECD framework provides insights from different disciplines, such as measurement, technology, and cognitive psychology, to help designers construct educational assessments of higher validity (Mislevy et al., 2003; Mislevy & Haertel, 2007). Even though the ECD framework defines multiple interconnected models, the current study focuses on three models of the ECD framework: the *student model*, the *task model*, and the *evidence model*.

The *student model* is about what is measured: defining one or more variables associated with the knowledge, skills, and competencies intended to be measured (Mislevy et al., 2003). These variables are therefore about students' characteristics. The *task model* in this study can be described as the task environment in which students demonstrate evidence for their reading comprehension skill (Mislevy & Haertel, 2007). This model focuses on where to measure the defined variable(s) from the student model. The task model also consists of a description of essential task features, also called task model variables.

Bridging the student model and the task model, is the *evidence model*. This model explains how to obtain evidence (i.e., values of observable variables, which are characteristics of

performance) from the behaviour of test takers in the context of their task (Mislevy et al., 1999). Moreover, the evidence model models the relationship between both the observable variables and the variables defined in the student model. More specifically, the measurement model describes how the observable variables are statistically dependent on the student model variables. The evidence model contains both *evidence rules* and a *measurement (or statistical) model* (Mislevy et al., 2003). *Evidence rules* are about describing observable variables, illustrating a test taker's performance on a certain task, with information on how these observations are scored (Mislevy et al., 1999). The *measurement model* offers information about the link between the student model variables and the observable variables, e.g., the measuring of the reading comprehension skill and the relation to the response on an item. In this study, the measurement model will help to make a connection between the target skill reading comprehension (student model), and the task model (specifically, the task features), by applying the psychometric parameters item difficulty and item discrimination. In this sense, this study contributes to reasoning with evidence about what task features of exam items bring about, by studying variations in item difficulty and item discrimination and relating these to task features.

Task Features Predicting Item Difficulty

Task features in Dutch language exams for reading comprehension and their predictive value for item difficulty have been investigated previously. This section will highlight the main findings of research into Dutch reading comprehension exams, on both reference level 2F and 3F.

2F Dutch comprehension exams

Roelofs, Keune, et al. (2021) studied predictive task features for item difficulty and item discrimination for 2F Dutch comprehension exams. In general, it was found that four groups of task features were predictive for item difficulty: text features, intrinsic task features, access skills, and item presentation features. Text features include the number of sentences or words in a text, the average sentences length and word length in a text, and the type token ratio: the proportion of unique words in a text to its overall word count). As a task feature on the text level, it was found that the *lexical richness of the text*, expressed by a so called type token ratio, increased item difficulty (Roelofs, Keune, et al., 2021). Roelofs, Keune, et al. (2021) mention another promising substantive text feature: propositional complexity. This measure reflects the density of the number of propositions (concepts and relationships) in a text. The present study includes several measures related to propositional complexity in the (text) task features. These measures are computed based on previously constructed concept maps, and further explained in the methods section.

Next to text features, intrinsic information processing task features, considered intrinsic to reading comprehension, were found to impact item difficulty. These intrinsic task features involved the scope of information to be used and the type of information task to be performed, both of which may elicit different amount of cognitive load for the reader. In descending order, the following intrinsic task features caused an increase in item difficulty: the *number of inferences to be formed*, the *assessment of propositions or arguments following the text*, the *amount of necessary information needed to answer an item concerning the entire text rather than individual passages*, the *organisation of information (elements) is central*, the item focuses on *determining rations between numbers*, the item focuses on *drawing a conclusion based on a combination of informative elements*, or the item focuses on *performing a task based on the (full) text or a text passage* (Roelofs, Keune, et al., 2021). Conversely, when an item focuses on *retrieving explicit information*, the item tend to be easier (Roelofs, Keune, et al., 2021).

The third group of task features involves the extent to which access skills are needed to reason correctly to the item. It was found that if *answering an item required an additional (access) skill,* the item became more difficult (Roelofs, Keune, et al., 2021). The authors noted that reliance on access skills, as these cause increased extraneous task load, should be avoided.
Finally, item presentation features impacted item difficulty. These involved features such as the degree to which the key represents the best possible response, whether or not eliminating response options is necessary, and the number of plausible response options. It was found the number of substantively plausible distractors contributed to an increased item difficulty.

Lieverse (2021) further investigated predictive task features of item difficulty and item discrimination for 2F Dutch reading comprehension exams. Building on Roelofs, Keune, et al. (2021), Lieverse (2021) included propositional complexity of the texts used. Several propositional text features were found to influence item difficulty: both the *number of centralised information elements* and the *type token ratio* increased item difficulty, whereas the *total number of information elements* caused a decrease in item difficulty. In addition, access skills tended to increase item difficulty. Lastly, the largest effect on item difficulty was found for inferences. An increase in the number of inferences to form in order to respond correctly to an item, causes an item to become more difficult.

3F Dutch comprehension exams

Roelofs, Postulart, et al. (2021) investigated the predictive value of task features for item difficulty and item discrimination in 3F Dutch comprehension exams. In their study, Roelofs, Postulart, et al. (2021) created concept maps to explore the relationship between information

elements present in a text on the one hand and the item difficulty and item discrimination on the other hand (propositional complexity). In doing so, they also distinguished both information elements and relationships, explicit or implicit in nature. More specifically, the implicit elements or relationships refer to elements or relationships that are not explicitly stated in the text but have to be inferred by the reader.

Roelofs, Postulart, et al. (2021) found several task features with a predictive value for item difficulty, both increasing or decreasing item difficulty. Regarding task features related to the concept maps, the *presence of the actor relationship type* decreased item difficulty. Additionally, they studied the impact of information task features as intrinsic features. In line with previous research in 2F Dutch reading comprehension exams, the *number of types of inferences to make* caused an increase in difficulty for the item. A third group of task features had the largest contribution to item difficulty, both positive and negative: the item presentation features. On the one hand, the *number of substantively plausible distractors* increased item difficulty, similar to research at reference level 2F (Roelofs, Keune, et al., 2021). On the other hand, the *degree to which the key is the best possible response*, indicates an easier item. I.e., the more accurate the key is, the easier the item is. Finally, in the group related to access skills, the feature *deducing meaning from punctuation marks*, increased item difficulty.

Overview of (groups of) task features predicting item difficulty

To summarise, in Table 1.1, an overview of the (groups of) task features found to predict item difficulty in the cited Dutch studies is given. Note: an increase in item difficulty is indicated by a plus (+), whereas a decrease in item difficulty is indicated by a minus (-).

Table 1.1

Overview of the (groups of) task features found to predict item difficulty

Groups of task features	2F, Roelofs, Keune, et al. (2021)	2F, Lieverse (2021)	3F, Roelofs, Postulart, et al. (2021)
Text features*	• type token ratio (+)	• number of centralised information elements (+) • type token ratio (+) • total number of elements (-)	• presence of the actor relationship type (-)
Intrinsic task features	• number of inferences to form (+) • assessment of propositions or arguments following the text (+) • amount of necessary information needed to answer an item concerning the entire text rather than individual passages (+) • organisation of information (elements) is central (+) • determining rations between numbers (+)	• number of inferences to form (+)	• number of types of inference to make (+)

	• drawing a conclusion based on a combination of informative elements (+) • retrieving explicit information (-)		
Access skills	• answering an item required an additional (access) skill (+)	• requires an additional access skill (+)	• deducing meaning from punctuation marks (+)
Item presentation features	• number of substantively plausible distractors (+)		• degree to which the key is the best possible response (-) • number of substantively plausible distractors (+)

*In the 2F report by Lieverse (2021) and the 3F report by Roelofs, Postulart, et al. (2021), also text features based on the concept maps were included

Task Features Predicting Item Discrimination

Task features in Dutch language exams for reading comprehension and their predictive value for item discrimination have been investigated previously. This section will highlight the main findings of research into Dutch comprehension exams, on both reference level 2F and 3F.

2F Dutch comprehension exams

Roelofs, Keune, et al. (2021) studied predictive task features for item difficulty and item discrimination for 2F Dutch comprehension exams. With regard to item discrimination, some predictive task features both increasing and decreasing the evidence strength were found (Roelofs, Keune, et al., 2021). Task features found were in the group of the intrinsic task features, the item presentation features, and the target skill. For the intrinsic task features, it was found that when there is a *follow-up task (or application task) based on textual information*, the evidence strength is increased. However, when there is *no direct information task, but a contemplative task transcending the text*, evidence strength decreased. Regarding the item presentation features, these played the strongest part in predicting item discrimination. On the one hand, it was found that the *degree to which the key is the best possible response*, indicates an item with higher evidence strength. On the other hand, when an item *requires eliminating response options*, evidence strength is decreased. For the group of the target skill, two features were found to impact item discrimination. First, when an item concerns the *ability to establish relationships between textual information and general knowledge, or building a situation model*, evidence strength increased. Second, when an item concerns the *ability to make relationships between text passages*, evidence strength decreased.

Lieverse (2021) further researched predictive task features of item difficulty and item discrimination for 2F Dutch reading comprehension exams, also including propositional complexity on a text level. Two text features related to the lexical richness of the text were found to influence item discrimination: the *number of the contrast relationship type* increased evidence strength, whereas the *number of the non-directly observable characteristics element* decreased evidence strength. Regarding the underlying information literacy, it was found an item focused on *retrieving*

social-communicative meaning decreased evidence strength. The main contribution to item discrimination was found for a certain *text topic*. Other significant contribution was from the group of item presentation features. The *degree to which the key is the best possible response*, indicates an item with higher evidence strength (similar to Roelofs, Keune, et al., 2021). Lastly, in the group of target skill, in line with Roelofs, Keune, et al. (2021), it was found that items concerning the ability *relating text passages and texts* had decreased evidence strength, as opposed to the ability *relating between textual information and common knowledge*, which increased the evidence strength.

3F Dutch comprehension exams

Roelofs, Postulart, et al. (2021) investigated the predictive value of task features for item difficulty and item discrimination in 3F Dutch comprehension exams, and created concept maps to further explore lexical richness of texts. Roelofs, Postulart, et al. (2021) found several task features with a predictive value for item discrimination, both increasing or decreasing item discrimination. Regarding the task features related to the concept maps, it was found the *degree to which relationships are implicit*, caused a decrease in item discrimination (i.e., evidence strength). A second group of task features were intrinsic task features. In this group, *matching a supporting argument with a point of view*, also caused a decrease in evidence strength.

Other predictive task features found had a positive influence on the evidence strength. The target skill feature *drawing conclusion based on (parts of) the text*, and the access skill feature *meaning of words is explicitly asked for*, both caused a moderate increase in evidence strength. The item presentation feature the *degree to which the key is the best possible response,* showed the biggest (positive) contribution to evidence strength.

Overview of (groups of) task features predicting item discrimination

To summarise, in Table 1.2, an overview of the (groups of) task features found to predict item difficulty is given. Note: an increase in item discrimination is indicated by a plus (+), whereas a decrease in item discrimination is indicated by a minus (-).

Table 1.2

Overview of the (groups of) task features found to predict item discrimination

Groups of task features	2F, Roelofs, Keune, et al. (2021)	2F, Lieverse (2021)	3F, Roelofs, Postulart, et al. (2021)
Text features*		• number of the contrast relationship type (+) • number of the non-directly observable characteristics element (-) • text topic (+)	• degree to which relationships are implicit (-)

Intrinsic task features	• follow-up task (or application task) based on textual information (+) • no direct information task, but a contemplative task transcending the text (-)	• retrieving social-communicative meaning (-)	• matching a supporting argument with a point of view (-)
Access skills		• relating text passages and texts (-) • relating between textual information and common knowledge (+)	• meaning of words is explicitly asked for (+)
Target skill	• ability to establish relationships between textual information and general knowledge, or building a situation model (+) • ability to make relationships between text passages (-)		• drawing conclusion based on (parts of) the text (+)
Item presentation features	• degree to which the key is the best possible response (+) • requires eliminating response options (-)	• degree to which the key is the best possible response (+)	• degree to which the key is the best possible response (+)

*In the 2F report by Lieverse (2021) and the 3F report by Roelofs, Postulart, et al. (2021), also text features based on the concept maps were included

The Use of Concept Maps to Represent the Situation Model

It was found in previous research that both the topic and content of a text induced differences in item difficulty (Roelofs, Keune, et al., 2021). Furthermore, Roelofs, Postulart, et al. (2021) explored the relationship between information elements and textual relationships and both item difficulty and item discrimination. The present study created data on additional task features, based on these concept maps. Specifically, per item, it will be coded which part(s) of the situation model, the mental representation that reflects the overall meaning of a text, as displayed by the concept maps, are needed in order to correctly respond to the item.

Concept maps are “graphical tools for organising and representing relationships between concept indicated by a connecting line linking two concepts” (Novak & Cañas, 2007, p. 29). Overall, a concept map is created keeping in mind the area of knowledge that is to be mapped, represented by a so-called focus question, after which key concepts are selected. More specifically, a concept is defined as “a perceived regularity (or pattern) in events or objects, or records of events or objects, designated by label” (Novak & Cañas, 2007, p. 33). The concepts are linked by a connecting line, where either linking words or phrases describe the relationship between two concepts. Two concepts with a connecting line describing the relationship between those two concepts is referred to as a proposition. In more detail, propositions are “statements about some object or event in the universe, either naturally occurring or constructed. Propositions contain two or more concepts connected using linking words or phrases to form a meaningful statement” (Novak & Cañas, 2006, p. 1). Moreover, concepts and propositions are mostly organised from more general, to more specific (so hierarchically) (Novak & Cañas, 2006, 2007).

Concept maps were initially developed to gain insight in changes in children's knowledge of science (Novak & Musonda, 1991). Concept maps helped the researchers to represent the children's knowledge, and therefore supported the identification of changes in children's understanding of science concepts. These concept maps were based on the assimilation theory in meaningful learning and retention processes by Ausubel (1963, 2000). He mentioned that within learner's cognitive structure, newly learned materials have to be related to concepts that are already existing in the cognitive structure (Ausubel, 1963, 2000). Relating the newly learned materials to existing concepts in one's cognitive structure, and the adapting of the newly learned materials to fit in one's cognitive structure is also called the principle of assimilation (Ausubel, 2000).

The present study considers previously created concept maps by Roelofs, Postulart, et al. (2021) as expert concept maps, and uses them as a proxy for an expert situation model. As concept maps represent a (personal) situation model, whether a concept map is a correct proxy of a situation model is an arbitrary decision. In the end, the concept map will be a mental representation of the overall meaning of the text, according to the creator of the concept map. However, as was suggested by Roelofs, Postulart, et al. (2021), it is expected that there might be concept maps (per text) that experts can agree upon, representing the area of knowledge that is to be mapped. I.e., it is expected that there are concept maps on which experts can agree that they represent the essence of the corresponding text.

Methods

Selected Items and Associated Exam Data

In the present study, two hundred items from eighteen texts originating from Dutch 3F reading comprehension exams were included. Out of an existing exam data set, pertaining response data for over twelve hundred items, an item sample of two hundred was drawn, following previous procedures described by Roelofs, Postulart, et al. (2021). The item sample was administered throughout various exam forms, and response data had been collected between 2015 and 2022.

In order to arrive at a representative sample of Dutch 3F reading comprehension items, a balanced selection was made representing both the original distribution of p-values and rir-values, and the representation of content categories, including text domain and text genre. Table 2 provides an overview of the texts included in the present study, with the corresponding number of items and item parameter values. The related item parameters used were pooled values as estimated in item analyses across various exam versions as administered throughout the indicated period.

Table 2

Overview of texts included in the present study, with their corresponding number of items and mean item parameter values

Text	Number of items	Mean p-value	Mean rir-value
A	12	63.0	23.6
B	12	49.2	12.8
C	12	57.3	11.8
D	11	71.2	17.6
E	12	56.2	7.6
F	9	67.1	17.6
G	12	58.6	13.1
H	12	67.2	14.2
I	12	71.6	17.2
J	12	60.9	12.4
K	12	75.5	16.1
L	12	67.7	21.8
M	9	63.6	17.6
N	9	83.9	22.2
O	9	62.0	20.3
P	12	50.9	8.5
Q	9	48.2	19.1
R	12	64.0	17.7
Overall	200	63.0	15.9

Regarding the representation on content categories, Table 3 provides an overview for both the exam and the present study. It was attempted to obtain somewhat similar proportions for the text domains in the present study, as compared to the proportions.

Table 3

Overview of the representation of the texts based on the text domains, for both the exam and the present study

	In exam		In study	
	N	**Proportion**	**N**	**Proportion**
1 – Political / legal	5	.15	3	.17
2 – Economy	22	.15	2	.11
3 – Social / societal	24	.40	8	.44
4 – Vital citizenship	7	.15	3	.17
5 – Career	3	.15	2	.11
Overall	61	1	18	1

Next, from the sample of two hundred items, several items (i.e., twelve) were removed because they had been neutralised by Cito for mall functioning or due to substantial errors. Subsequently, six items were removed from the data file because these involved a small group of items with quite a different response format, in this case, matrix items, involving multiple statements to be rated in one item. After removal of these items, 182 items remained in the sample. Table 4 provides a full overview of the texts included, with per text the number of neutralised items, matrix items, the item parameter values, and text domain. Regarding the item parameters, the reduction in item sample caused the mean p-value of all items to increase from 63.0 to 63.2; the mean rir-value of all items increased from 15.9 to 16.5. The rir-value will be reported on a scale from 0-100, as is the standard at Cito. With regards to the p-value, the notation on a scale of 0-100 will be used throughout the rest of this study.

Table 4

Overview of included texts, with their corresponding number of items, number of neutralised items, number of matrix items, item parameters, and text domain

Text	Number of items	Number of neutralised items	Number of matrix items	Mean p-value	Mean rir-value	Text domain
A	12	0	0	63.0	23.6	1 – political / legal
B	12	0	0	49.2	12.8	5 – career
C	12	0	0	57.3	11.8	5 – career
D	10	0	1	72.5	17.6	4 – vital citizenship
E	11	1	0	57.6	8.8	3 – social / societal
F	8	1	0	63.0	17.6	2 – economy
G	11	0	1	62.9	13.8	4 – vital citizenship
H	10	2	0	67.0	17.0	3 – social / societal
I	9	3	0	72.2	18.9	3 – social / societal
J	12	0	0	60.9	12.4	4 – vital citizenship
K	11	1	0	73.3	16.0	1 – political / legal
L	12	0	0	67.7	21.8	1 – political / legal
M	8	1	0	59.0	17.6	2 – economy
N	8	0	1	85.3	20.9	3 – social / societal
O	8	0	1	60.1	20.1	3 – social / societal
P	9	3	0	56.6	13.7	3 – social / societal
Q	8	0	1	45.3	19.0	3 – social / societal
R	11	0	1	66.9	17.6	3 – social / societal
Overall	182	12	6	63.2	16.5	-

Coding task features

In order to investigate the contribution of the task features to item difficulty and item discrimination of coding scheme for item features was developed based on Roelofs, Postulart, et al. (2021). This coding scheme organised as a form per item, provided a comprehensive overview of all task features involved with a certain item, as well as the p-value and rir-value. In Annex 2, an example of such a form in Microsoft Access is presented. For the present study, in total, twelve groups of task features were distinguished, mostly based on Roelofs, Postulart, et al. (2021). These were groups of features regarding: proposition, type of relationships, text content, information task, type of inferences, number of inferences, meta task, implicit language task, explicit language task, presentation, and target skill. Regarding the task features, a total of 58 features were included, divided over twelve groups (see Table 5). The present study mainly focused on creating extra task features based on the previously constructed concept maps by Roelofs, Postulart, et al. (2021). In total, 26 features out of the total of 58 features were created and scored in the present study (45%). These extra features were concentrated around coding propositions needed by test takers to arrive at the correct response. This will be elaborated upon in the upcoming section.

Table 5

Overview of the groups of task features and the number of features included

Groups of task features	Number of features in group	Adopted from Roelofs, Postulart, et al. (2021)
Proposition	5	No
Type of relationships	21	No
Text content – Text	1	Yes
Text content – Domain	1	Yes
Information task	9	Yes
Types of inferences	4	Yes
Number of inferences	1	Yes
Meta task	3	Yes
Implicit language task	3	Yes
Explicit language task	2	Yes
Presentation features	7	Yes
Target skill	1	Yes
Total	58	-

Table 6 further elaborates on the groups of task features, and also lists the item features included per group. The following sections first focus on the features regarding the proposition and the type of relationships, and more specifically on the coding process of these propositions (deemed necessary to arrive at a correct response to the item). The focus then shifts to the coding scheme for task features as developed by Roelofs, Postulart, et al. (2021).

Table 6

An overview of all item features per group

Group	Item features
Proposition	• Number of propositions needed in order to correctly answer the item • Number of implicit information elements involved with correctly answering the item • Number of implicit relationships between information elements involved with correctly answering the item • Total number of both implicit information elements and relationships involved with correctly answering the item • Number of different types of relationships involved with correctly answering the item
Type of relationships	• Kind/type • Characteristic/feature • Actor • Effect • Causal • Temporal

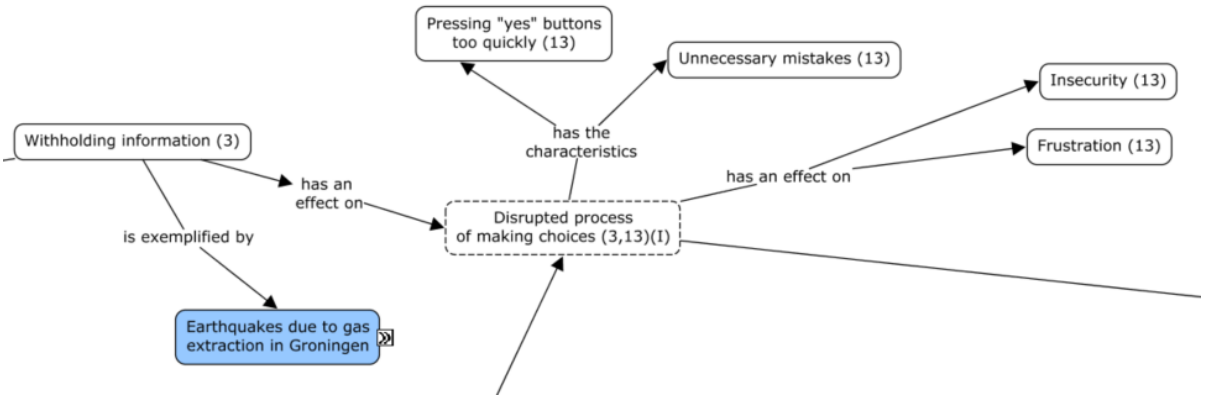
	• Example/illustration • Concluding • Evaluative • Means-goal • Motive • Explanatory • Location • Source (of information) • Recipient • With respect to • Contrasting • Condition • If-then • Implying • Others •
Text content - Text	• Name of the corresponding text
Text content - Domain	• Text domain Dutch vocational education
Information task	• Finding justification • Similarities and differences • Matching a point of view with a supporting argument • Separating relevant information • Searching for literal information • Meta task • Selecting a written summarising sentence • Finding arguments (including advantages and disadvantages) • Drawing a conclusion based on a combination of information elements
Type of inferences	• Connecting information • Filling an information gap with factual knowledge • Deriving new information by logic reasoning • Finding a subordinate label
Number of inferences	• Number of inferences
Meta task	• Argumentation theory • Text structure • Social communicative meaning making
Implicit language task	• Meaning of words • Expression(s) • Derive meaning of punctuation marks
Explicit language task	• Meaning of words • Expression(s)
Presentation features	• Key is the best possible response • Eliminating response options is necessary • Number of plausible response options (plausible distractors) • Overlap between key and textual information mentioned • Overlap distractors with textual information mentioned • Overlap between item stem and the key • Overlap between item stem and distractors
Target skill	• Cluster

Coding propositions deemed necessary to arrive at a correct response to the item

As was found in previous studies (e.g., Lumley et al., 2012), students have to either identify or infer some necessary information based on the mental representation of the text in order to arrive at a correct response on a comprehension item. In the present study, previously created expert concept maps by Roelofs, Postulart, et al. (2021) were considered as a plausible representation of the text content. These expert concept maps exist of networks of propositions each consisting of concept 1, concept 2 and the connecting relationship between these concepts. Some of the concepts and the relationships between them are not literally stated within the text but needed to be inferred (i.e., implicit concepts or relationships). In drawing the concepts, using the program CmapsTools, an implicit concept or relationship was expressed by using dashed squares for the concepts and dashed lines for the relationship. Moreover, implicit concepts or relationships included "(I)". When it was applicable, also the section number was noted in the concept or relationship space (see Figure 1 for an example).

Figure 1

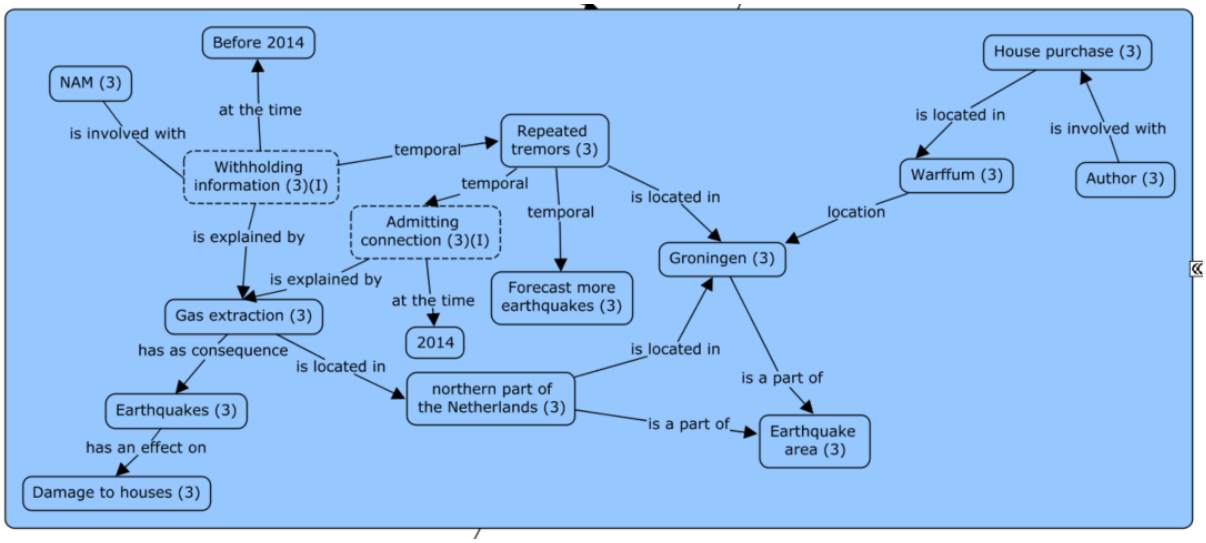
A small part of the expert concept map of one of the texts (Text O)



Furthermore, some overarching concepts were indicated in a blue colour, and could be unfolded to display the complete network (see Figure 2). A blue coloured concept can be seen as a higher order concept, which includes a small network of propositions and relationships.

Figure 2

The concept "Earthquakes due to gas extraction in Groningen", as elaborated upon in the expert concept map of one of the texts (Text O)



Before coding the propositions deemed necessary to arrive at a correct response to the item, the expert concept maps were thoroughly checked, and some additions were made to some expert concept maps to make them better suit the other concept maps. I.e., it was observed that in some expert concept maps, the concepts used were more overarching concepts, on a higher hierarchical level. These concepts were then split into several concepts, so they were more in line with the other expert concept maps.

Next, for each of the comprehension items, it was decided which of the propositions in the concept map of the text we deemed necessary to be used in order to arrive at a correct answer to the (reading comprehension) item. The coding of the applicability of a proposition required three preparing steps. First, the written content of the proposition components, i.e., concept 1, concept 2 and the connecting relationship between the concepts – was exported to a database, whereby each proposition and its components appeared in one data row. Second, the proposition components were coded as 'implicit' when these had to be inferred by the expert composer of the concept maps and as 'explicit' when the component could be literally drawn for the text. Third, the connecting relationship was categorised, using brief descriptions (such as 'causal', 'means-goal', 'temporal', 'concluding'). The categorisation of the connecting relationship will be discussed later in this paragraph.

Subsequently, during the actual coding of the necessary information, for each item, the propositions deemed necessary in order to arrive at an answer to the question were coded (0 'not

applicable', 1 'applicable', respectively). The coding activities yielded data for eighteen texts, where the proposition is the unit of observation, regarding the following variables:

- the written content of concept 1;
- the written content of concept 2;
- the written content of the relationship between concept 1 and concept 2;
- the implicitness of concept 1, concept 2, and their relationship (0 = explicit; 1 = implicit);
- the type of relationship within the proposition;
- the necessity to use the proposition for each of the items in order to arrive at a correct response to the comprehension question (0, 1);
- the title of the text.

After finishing this process for all items in all texts, some calculations were done to create new task features related to the concept maps on the item level. These new features first elaborated on below and are summarised in Table 7.1.

- the *number of propositions needed in order to correctly answer the item* (which is by definition the same as the number of relationships needed to arrive at the correct answer), which was calculated by summing the scores given for necessity to use the proposition to arrive at a correct response;
- the *number of implicit information elements involved with correctly answering the item*, which was calculated by summing the implicitness scores given for all information elements in the text for a specific item;
- the *number of implicit relationships between information elements involved with correctly answering the item*, which was calculated by summing the implicitness scores given for all relationships in the text for a specific item;
- the *total number of both implicit information elements and relationships involved with correctly answering the item,* which was calculated by summing both the *number of implicit information elements involved with correctly answering the item*, and the *number of implicit relationships between information elements involved with correctly answering the item*;
- the *number of different types of relationships involved with correctly answering the item*, which was calculated by counting the various types of relationships in the concept maps, that test takers are expected to use in order to arrive at a correct response. The different types of relationships are summarised in Table 7.2.

Table 7.1

An overview of the group 'proposition' features, with their scoring

Proposition features	Score
• Number of propositions needed in order to correctly answer the item	• Number (0 – N)
• Number of implicit information elements involved with correctly answering the item	• Number (0 – N)
• Number of implicit relationships between information elements involved with correctly answering the item	• Number (0 – N)
• Total number of both implicit information elements and relationships involved with correctly answering the item	• Number (0 – N)
• Number of different types of relationships involved with correctly answering the item	• Number (0 – N)

Regarding the categorisation of the connecting relationship between concepts, in total 21 types of relationships have been coded. For a complete overview of all relationships coded, see Table 7.2. The description of the relationships was (partially) adopted from Roelofs, Postulart, et al. (2021). The relationships were separately coded, based on the type of relationship from the previously coded proposition.

Table 7.2

An overview of the group 'type of relationships' features

Type of relationship features	Score*
• Kind/type: a breakdown by types of information element A (e.g., A1, A2), that are common under A. Individual types may show more or less relatedness to A.	• 0, 1
• Characteristic/feature: defining and distinguishing characteristics or features. The relationship between an information element and its properties, both concrete and abstract.	• 0, 1
• Actor: one information element (human) is an actor for another information element.	• 0, 1
• Effect: one information element has an effect on another; either negative, positive, or neutral.	• 0, 1 • 0, 1
• Causal: there is a causal relationship between information elements.	• 0, 1
• Temporal: information elements are ordered in time order, such as a historical period or comparing past and present.	• 0, 1
• Example/illustration: the information element is an example for clarification or illustration of another information element (animal: dog).	• 0, 1
• Concluding: the first information element leads to a conclusion in the second information element.	• 0, 1
• Evaluative: one information element is an actor and has an evaluative expression towards another information element.	• 0, 1 • 0, 1

• Means-goal: similar to causality, there is a temporal sequence for this type of relationship. Means come before the goal to be achieved. Arises from a human motive.	• 0, 1
• Motive: the first information element motivates to realise the second information element.	• 0, 1
• Explanatory: one information element offers an explanation or elaboration of another information element.	• 0, 1
• Location: one information element refers to the (physical) location of another information element.	• 0, 1
• Source (of information): one information element is the source of information in another information element.	• 0, 1
• Recipient: one information element (human) is the recipient for another information element.	• 0, 1
• With respect to: one information element takes into account another information element to a certain extent.	• 0, 1
• Contrasting: information elements are contradicting each other, in relation to information in the text and its understanding.	• 0, 1
• Condition: one relationship is conditional for another, e.g., information element B cannot be there without (conditional) information element A.	• 0, 1
• If-then: similar to causality, there is a temporal sequence for this type of relationship. 'If' happens before 'then'. Does not arise from a human motive.	• 0, 1 • 0, 1
• Implying: one information element implies another information element.	
• Others: other types of relationships that could not be classified in the twenty types mentioned above are collected as 'others'.	

* 0: not applicable, 1: applicable

The proposition data were joined to the item parameter data using the text and item number as key variables. In such a way, further analyses could be performed in R, as described in results section.

Coding scheme for task features

A third group of features included text content features, which were scored on text level. I.e., all items belonging to a certain text obtained a similar score. As items are nested in the text they belong to, and texts are different (e.g., more or less difficult), the name of the corresponding text was included as a feature. Moreover, the text domain of Dutch vocational education in which a text has been classified, scored on the text level, was included. An overview of these text content features, and their scoring is given in Table 7.3.

Table 7.3

An overview of the group 'text content' features, with their scoring

Text content features	**Score**
• Name of the corresponding text	• Text A, Text B, Text C, Text D, Text E, Text F, Text G, Text H, Text I, Text J, Text K, Text L, Text M, Text N, Text O, Text P, Text Q, Text R
• Text domain Dutch vocational education	• 1 – political / legal; 2 – economy; 3 – social / societal; 4 – vital citizenship; 5 – career

Another group of features included, regarding the information task, focus on the mental action a test taker has to perform, based on the information from the text. This mental action will support forming a mental representation of the information from the text. Based on both theory and previous research into 2F Dutch reading comprehension exams (Roelofs, Keune, et al., 2021), it is mentioned a distinction is made between *supporting information tasks*, which focus on processing different information elements to form micro propositions, and *integrative information tasks*, which focus on building macro propositions (Kintsch & Rawson, 2005). The information task features, either supporting or integrative, are summarised in Table 7.4.

Table 7.4

An overview of the group 'information task' features, with their scoring

Information task features	**Score***
Supporting information tasks (micro propositions)	
• Finding justification	• 0, 1
• Similarities and differences	• 0, 1
• Matching a point of view with a supporting argument	• 0, 1
• Separating relevant information	• 0, 1
• Searching for literal information	• 0, 1
Integrative information tasks (macro propositions)	
• Meta task	• 0, 1
• Selecting a written summarising sentence	• 0, 1
• Finding arguments (including advantages and disadvantages)	• 0, 1
• Drawing a conclusion based on a combination of information elements	• 0, 1

* 0: not applicable, 1: applicable

Two other groups of features, the type and number of inferences, also focus on the mental action a test taker has to perform, to support forming a mental representation of the information from the text. It is mentioned previous research into 2F Dutch reading comprehension exams found that both the type and number of inferences used in processing and applying necessary text information for answering an item, strongly influences item difficulty (Roelofs, Keune, et al., 2021). Regarding the type of inferences, it is mentioned a division of four was created: connecting information, filling an information gap with factual knowledge, deriving new information by logic reasoning, and finding a subordinate label. Kispal (2008) elaborates on these types of inferences.

Connecting information. These types of inferences are about connections to be made between different sentences, and often include referencing words. It is mentioned that this type of inference

is only scored when the inference was of importance for understanding the text section that was involved with the item, as this type of inference frequently occurs in all texts. *Filling an information gap with factual knowledge*. This type of inference focuses on filling the missing or only implicitly mentioned information gaps. Factual knowledge should help a reader to find the connection between sentences. *Deriving new information by logic reasoning*. This type of inference also focuses on filling in missing or only implicitly mentioned information gaps, but by using logic reasoning. The readers have to use information elements from the texts to derive new information. *Finding a subordinate label*. These types of inference entail the coherent representation of the text. It is about finding the overarching ideas regarding the theme, main idea or moral of the text. The reader should derive this subordinate label by local pieces of text information. The task features regarding type of inferences, and number of inferences, and their scoring can be found in respectively Table 7.5 and Table 7.6.

Table 7.5

An overview of the group 'type of inferences' features, with their scoring

Type of inference features	Score*
• Connecting information	• 0, 1
• Filling an information gap with factual knowledge	• 0, 1
• Deriving new information by logic reasoning	• 0, 1
• Finding a subordinate label	• 0, 1

* 0: not applicable, 1: applicable

Table 7.6

An overview of the group including a 'number of inferences' feature, with the scoring

Number of inference feature	Score
• Number of inferences	• Number (0 – N)

As a different group of features, Roelofs, Postulart, et al. (2021) included meta task. Meta task still focuses on the mental action a test taker has to perform, to support forming a mental representation of the information from the text. It is mentioned that compared to the previous

groups of features, where the processing of information that focuses more on direct manipulation of information, meta tasks are transcending texts. In other words, a test taker has to reflect on the text using knowledge from outside the text. Roelofs, Postulart, et al. (2021) made a distinction for three different types of meta tasks: argumentation theory, text structure, and social communicative meaning making. *Argumentation theory.* It is mentioned that as items involving argumentation and the application of concepts from argumentation theory have a meta-character. *Text structure.* Regarding text structure, it is mentioned items can (implicitly) put test takers in the role of editor. The item e.g., asks about the function of a certain ordering in the text (i.e., text structure). *Social communicative meaning making.* As a last type of meta task, the reflection on social communicative context mentioned. This context involves not only characteristics of a genre, but also participants in conversation and their background. This requires a test taker to have knowledge about these types of characteristics. An overview of these features and their scoring is given in Table 7.7.

Table 7.7

An overview of the group 'meta task' features, with their scoring

Meta task features	**Score***
• Argumentation theory	• 0, 1
• Text structure	• 0, 1
• Social communicative meaning making	• 0, 1

* 0: not applicable, 1: applicable

Roelofs, Postulart, et al. (2021) mention another group of task features: the conditional language task. The language task is conditional in the sense that it requires test takers to have knowledge of certain word meanings and understanding of certain expressions and punctuation marks. In the present study, the conditional language task is split into implicit language task, and explicit language task. An overview of the features included can be found in Table 7.8 and Table 7.9. These features are scored when directly mentioned in an item, or (implicitly) needed in finding the information needed in order to correctly answer the item. It is mentioned the conditional language task is comparable to the verbal part of the skill foci as mentioned by Deane et al. (2011).

Table 7.8

An overview of the group 'implicit language task' features, with their scoring

Implicit language task features	Score*
• Meaning of words	• 0, 1
• Expression(s)	• 0, 1
• Derive meaning of punctuation marks	• 0, 1

* 0: not applicable, 1: applicable

Table 7.9

An overview of the group 'explicit language task' features, with their scoring

Explicit language task features	Score*
• Meaning of words	• 0, 1
• Expression(s)	• 0, 1

* 0: not applicable, 1: applicable

Subsequently, based on Roelofs, Postulart, et al. (2021) a category for extrinsic task features was used. These extrinsic task features do not cover the target skill but do influence a test takers' performance. These extrinsic task features, or features regarding the presentation of the item, increase the extraneous cognitive task load (Sweller, 2010). *Features of task presentation.* The extraneous cognitive task load might be increased by (but not limited to) a key that does not fully cover the correct response, causing confusion for test takers about the correct answer. Another feature that might impact the extraneous cognitive task load is an unfocused or unclear question (stem) that can only be understood using the response options, causing the need for eliminating response options. A final task presentation feature, involved the number of plausible response options, or plausible distractors. This feature does not necessarily elicit extraneous cognitive task load, but the presence of different plausible response options causes a test taker to be very specific in their answer. *Lexical overlap*. Furthermore, lexical overlap was coded: between key and distractors on the one hand, and textual information on the other hand. Another type of lexical overlap was coded between the item stem and both the key and the other response options. Both types can affect the item difficulty (Roelofs, Postulart, et al., 2021). Presentation features are summarised with their scoring in Table 7.10.

Table 7.10

An overview of the group 'presentation features', with their scoring

Presentation features	Score
Features of task presentation	
• Key is the best possible response	• Not the best possible response: 0%; partially best possible response: 50%; best possible response: 100%
• Eliminating response options is necessary	• Not necessary: 0; partially necessary: 0,5; necessary: 1
• Number of plausible response options (plausible distractors)	• 0-1-2-3
Lexical overlap	
• Overlap between key and textual information mentioned	• 0, 1*
• Overlap distractors with textual information mentioned	• 0, 1*
• Overlap between item stem and the key	• 0, 1*
• Overlap between item stem and distractors	• 0, 1*

* 0: not applicable, 1: applicable

A final group, including one feature, concerned the target skill as defined by the exam constructors. These target skills were based on the target skills as defined by the reference level 3F (see Annex 1). Also, it was mentioned argumentation was added to these target skills. Furthermore, it was mentioned the scoring for the target skill feature was based on the ones named in the syllabus (College voor Toetsen en Examens, 2017). An overview can be found in Table 7.11.

Table 7.11

An overview of the group including a 'target skill' feature, with the scoring

Target skill feature	Score
• Cluster	• 1 – goal / type of text; 2 – text structure; 3 – main and side issues / summarising; 4 – understanding and interpretation of information; 5 – argumentation

Results

In this section the results of this study are reported: both descriptive analyses and analyses aimed at answering the research question. Firstly, descriptive statistics regarding the item parameters and the representation of task features of the studied item set are given. Secondly, the associations between (groups) of task features and item parameters are explored using correlation analyses. For the categorical task features, analyses of variance were performed, and in addition, eta squared was calculated to check the magnitude of the relations. Thirdly, the predictive value of text regarding the item parameters is described. Fourthly, results from multiple regression are presented, for both predicting item difficulty, and item discrimination.

Descriptive Statistics: Item Parameters and Task Features

Table 8 shows descriptive statistics for the item parameters in the item sample. The item difficulty amounted from a minimum value of 16.0 to a maximum value of 93.0, with a mean of 63.2. This mean corresponds with items of moderate difficulty. With regard to the item discrimination, a minimum value of -5.0 was found, and a maximum value of 40.0, with a mean of 16.5. This is a relatively low value for item discrimination and indicates thin evidence. I.e., items do not discriminate well between students, taking into account practical rules of thumb for this index (Zijlmans et al., 2018). An overview of the mean p-value and mean rir-value per text can be found in Table 9.

Table 8

Descriptive statistics for item parameters in the item sample: minimum, maximum, mean, median and interquartile range (IQR)

Statistic	Min	Max	Mean	Median	IQR
p-value	16.0	93.0	63.2	67.0	[50.0;76.0]
rir-value	-5.0	40.0	16.5	17.0	[12.0;21.0]

An overview of the different groups of task features, with the number of features in the group is provided in Table 5.

Prevalence of proposition features

In the following paragraphs, the added task features by the present study will be highlighted and their prevalence in the exam items will be discussed. These task features relate to the propositions, type of relationships and (implicit) elements coded by the author, based on the concept maps constructed previously by Roelofs, Postulart, et al. (2021). The mean of these task features, per text are given in Table 9. Other information provided in the table pertains to the number of items, mean p-value and mean rir-value, for each text.

Table 9

Overview of number of items, and average p/rir-value per text, and proposition task features

				Features of necessary information needed to arrive at a correct response (yielded by analysis of item-concept map coverage)				
Text	**Number of items**	**Mean p-value**	**Mean rir-value**	**Mean N propositions**	**Mean N implicit information elements**	**Mean N implicit relation-ships**	**Score implicitness of information (elements + relationships)**	**Mean N different relationship types**
A	12	63.0	23.6	18.7	6.9	2.1	9.0	6.7
B	12	49.2	12.8	8.9	2.2	1.2	3.3	4.9
C	12	57.3	11.8	12.7	5.2	0.1	5.3	5.3
D	10	72.5	17.6	9.9	1.3	0.0	1.3	3.7
E	11	57.6	8.8	7.6	2.7	0.0	2.7	4.5
F	8	63.0	17.6	13.9	1.9	0.0	1.9	5.4
G	11	62.9	13.8	10.9	2.0	0.0	2.0	4.4
H	10	67.0	17.0	11.0	2.1	0.0	2.1	5.0
I	9	72.2	18.9	9.1	0.4	0.0	0.4	5.1
J	12	60.9	12.4	9.4	3.7	0.0	3.7	4.9
K	11	73.3	16.0	6.6	0.5	0.0	0.5	4.2
L	12	67.7	21.8	9.8	2.4	0.7	3.1	4.3
M	8	59.0	17.6	8.0	1.6	0.0	1.6	4.9
N	8	85.3	20.9	9.1	1.9	0.0	1.9	3.6
O	8	60.1	20.1	10.4	2.0	0.0	2.0	4.1
P	9	56.6	13.7	10.2	2.2	0.9	3.1	3.3
Q	8	45.3	19.0	5.6	1.5	0.0	1.5	3.4
R	11	66.9	17.6	9.7	0.7	0.0	0.7	5.2
Overall	182	63.2	16.5	10.2	2.4	0.3	2.7	4.7

The features listed in the columns of Table 9, next to text, number of items, and mean p-value and mean rir-value, pertain to features of necessary information needed to arrive at a correct response, found after analysis of the concept map coverage. With regard to the mean p-value, some texts possessed a lower mean p-value, indicating relatively difficult items (text B and text Q, resp. 49.2 and 45.3). Moreover, some texts possessed a higher mean p-value, indicating relatively easy items (texts D, I, K and N, resp. 72.5, 72.2, 73.3 and 85.3). Considering the mean rir-value, none of the texts indicated thick evidence (rir-value >.3). However, one text had a remarkably low rir-value of 8.8 (text E), indicating very weak evidence. Three of the texts, text A, N and O, possessed a mean rir-value which indicated emerging evidence (resp. 23.6, 20.9 and 20.1).

The table also yields a picture regarding the features of necessary information needed to arrive at a correct response, for which four measures were used. First, the mean *number of propositions needed in order to correctly answer the item* ranged from 5.6 to 18.7 (resp. text Q and text A). An interesting note is that text A, possessing the highest mean number of propositions (18.7), also possesses the highest mean rir-value (23.6). Needing a high number of propositions might be an

indicator for items with more evidence strength. Moreover, text Q, with the lowest mean number of propositions (5.6), has a relatively low mean p-value (45.3).

Second, regarding the implicitness of necessary information Table 9 provides additional information. Therefore, the total mean *number of both implicit information elements and relationships needed to correctly respond to the item* was calculated, by summing the mean *number of implicit information elements* and the mean *number of implicit relationships between information elements*. Consequently, the results for the composing parts follow a similar trend. On all indicators of implicitness of necessary information, the item set belonging to text A showed the highest value, with the highest degree of implicitness, 9.0, the highest mean number of necessary implicit information elements, 6.9, and the highest mean number of implicit relationships (2.1). It could be that implicitness, either in information elements, relationships, or both, is an indication for a higher rir-value (text A: 23.6). Similarly, text I and K showed a rather low value on the degree of implicitness (respectively 0.4 and 0.5), which was based on their rather low mean number of necessary implicit information elements (respectively 0.4 and 0.5), and a score of 0 for the mean number of implicit relationships. It is noteworthy that these texts both have a relatively high p-value (resp. 72.2 and 73.3).

A third measure was composed, the mean *number of different types of relationships involved with correctly answering the item*. For instance, cause-effect, temporal order, among the necessary information elements (see Table 7.1 in method section for a complete overview). Text A shows the highest mean number of different relationships (6.7). It is noteworthy that text A had a relatively high rir-value of 23.6. Moreover, a much lower mean number of different relationship types required to arrive at the correct response was found for text D, N, P and Q (resp. 3.7, 3.6, 3.3 and 3.4).

Prevalence of types of relationships

Moreover, proposition task features included the type of relationship found in the proposition. An overview of the prevalence of these different relationship types can be found in Table 10.1 and Table 10.2. The relationships that were found most frequently across texts, in descending order, are *characteristic / feature* (N = 333), *effect* (N = 311) and *kind / type* (N = 261). The relationships found the least across texts, in ascending order, are *implying* (N = 12), *with respect to* (N = 18) and *condition* (N = 19).

Table 10.1

Overview of the prevalence of task features in the group type of relationship (1/2)

				Type of relationship																			
Text	**Number of items**	**Mean p-value**	**Mean rir-value**	**Kind / type**		**Characteristic / feature**		**Actor**		**Effect**		**Causal**		**Temporal**		**Example / illustration**		**Concluding**		**Evaluative**		**Means - goal**	
				N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop
A	12	63.0	23.6	12	.05	16	.05	4	.03	9	.03	1	.01	0	.00	4	.03	1	.03	0	0	0	.00
B	12	49.2	12.8	31	.12	13	.04	7	.05	7	.02	11	.16	2	.05	3	.03	0	.00	1	.02	0	.00
C	12	57.3	11.8	38	.15	14	.04	5	.03	10	.03	12	.17	6	.15	15	.13	0	.00	3	.06	3	.03
D	10	72.5	17.6	18	.07	14	.04	7	.05	22	.07	1	.01	6	.15	6	.05	1	.03	0	0	8	.09
E	11	57.6	8.8	13	.05	24	.07	8	.05	29	.09	2	.03	6	.15	8	.07	2	.06	4	.08	6	.07
F	8	63.0	17.6	10	.04	28	.08	10	.07	15	.05	3	.04	3	.08	12	.10	3	.09	1	.02	1	.01
G	11	62.9	13.8	10	.04	5	.02	5	.03	6	.02	19	.28	0	.00	7	.06	2	.06	3	.06	18	.20
H	10	67.0	17.0	4	.02	24	.07	4	.03	36	.12	3	.04	0	.00	0	.00	8	.24	3	.06	12	.14
I	9	72.2	18.9	10	.04	24	.07	10	.07	17	.05	0	.00	1	.03	0	.00	4	.12	5	.1	0	.00
J	12	60.9	12.4	16	.06	19	.06	1	.01	42	.14	0	.00	0	.00	8	.07	0	.00	3	.06	0	.00
K	11	73.3	16.0	4	.02	16	.05	8	.05	6	.02	0	.00	0	.00	3	.03	1	.03	2	.04	2	.02
L	12	67.7	21.8	23	.09	30	.09	25	.17	36	.12	9	.13	6	.15	8	.07	0	.00	8	.16	11	.13
M	8	59.0	17.6	5	.02	19	.06	5	.03	16	.05	1	.01	0	.00	9	.08	2	.06	5	.1	1	.01
N	8	85.3	20.9	4	.02	18	.05	8	.05	7	.02	2	.03	0	.00	2	.02	6	.18	1	.02	5	.06
O	8	60.1	20.1	11	.04	17	.05	5	.03	10	.03	1	.01	0	.00	1	.01	3	.09	6	.12	4	.05
P	9	56.6	13.7	16	.06	21	.06	9	.06	16	.05	2	.03	6	.15	10	.08	1	.03	1	.02	2	.02
Q	8	45.3	19.0	6	.02	2	.01	11	.07	9	.03	0	.00	4	.10	11	.09	0	.00	0	0	5	.06
R	11	66.9	17.6	30	.11	29	.09	19	.13	18	.06	2	.03	0	.00	13	.11	0	.00	4	.08	10	.11
Overall	182	63.2	16.5	261	1	333	1	151	1	311	1	69	1	40	1	120	1	34	1	50	1	88	1

Table 10.2

Overview of the prevalence of task features in the group type of relationship (2/2)

				Type of relationship																					
Text	**Number of items**	**Mean p-value**	**Mean rir-value**	**Motive**		**Explanatory**		**Location**		**Source (of information)**		**Recipient**		**With respect to**		**Contrasting**		**Condition**		**If - then**		**Implying**		**Others**	
				N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop	N	Prop
A	12	63.0	23.6	2	.07	5	.03	4	.08	2	.07	0	0	0	.00	1	.05	1	.05	0	.00	0	.00	7	.16
B	12	49.2	12.8	3	.11	14	.09	1	.02	0	.00	2	.08	0	.00	1	.05	3	.16	0	.00	0	.00	9	.20
C	12	57.3	11.8	0	.00	10	.06	2	.04	3	.10	2	.08	2	.11	0	.00	0	.00	0	.00	0	.00	4	.09
D	10	72.5	17.6	0	.00	6	.04	0	.00	1	.03	3	.12	1	.06	1	.05	0	.00	1	.03	0	.00	0	.00
E	11	57.6	8.8	0	.00	17	.11	1	.02	0	.00	0	0	0	.00	4	.18	0	.00	2	.06	0	.00	0	.00
F	8	63.0	17.6	0	.00	16	.10	7	.13	1	.03	0	0	2	.11	0	.00	1	.05	0	.00	1	.08	0	.00
G	11	62.9	13.8	0	.00	13	.08	6	.12	1	.03	0	0	0	.00	1	.05	4	.21	0	.00	1	.08	1	.02
H	10	67.0	17.0	0	.00	8	.05	3	.06	5	.17	3	.12	1	.06	3	.14	3	.16	2	.06	0	.00	0	.00
I	9	72.2	18.9	0	.00	4	.03	0	.00	4	.14	1	.04	1	.06	0	.00	0	.00	10	.29	0	.00	0	.00
J	12	60.9	12.4	12	.43	3	.02	3	.06	6	.21	3	.12	0	.00	0	.00	2	.11	0	.00	0	.00	0	.00
K	11	73.3	16.0	0	.00	6	.04	0	.00	1	.03	2	.08	3	.17	0	.00	0	.00	0	.00	0	.00	0	.00
L	12	67.7	21.8	0	.00	15	.10	5	.10	0	.00	0	0	4	.22	3	.14	0	.00	16	.46	1	.08	10	.22
M	8	59.0	17.6	3	.11	9	.06	10	.19	0	.00	6	.24	1	.06	1	.05	0	.00	0	.00	1	.08	2	.04
N	8	85.3	20.9	0	.00	7	.05	0	.00	1	.03	1	.04	1	.06	0	.00	0	.00	1	.03	5	.42	2	.04
O	8	60.1	20.1	0	.00	4	.03	0	.00	0	.00	1	.04	1	.06	2	.09	2	.11	0	.00	1	.08	2	.04
P	9	56.6	13.7	2	.07	6	.04	10	.19	1	.03	0	0	1	.06	0	.00	2	.11	1	.03	2	.17	2	.04
Q	8	45.3	19.0	5	.18	9	.06	0	.00	2	.07	1	.04	0	.00	1	.05	0	.00	2	.06	0	.00	0	.00
R	11	66.9	17.6	1	.04	2	.01	0	.00	1	.03	0	0	0	.00	4	.18	1	.05	0	.00	0	.00	6	.13
Overall	182	63.2	16.5	28	1	154	1	52	1	29	1	25	1	18	1	22	1	19	1	35	1	12	1	45	1

Further notable features are the relative large proportion in *causal* relationships in text G (.28), the relative large proportion in *motive* relationships in text J (.43), the relative large proportion of *if-then* relationships in both text I and L (respectively .29 and .46), the relative large proportion of *implying* relationships in text N, and the relative large proportion in *other* relationships in text L (.22). For the *motive* relationship, the *if-then* relationship and the *implying* relationship, it should be mentioned these types of relationships were found in less than half of the eighteen texts. This could explain why the proportions for some other texts are relatively large.

Correlation Analyses and Effect Sizes: Item Parameters and Task Features

Pearson product moment correlation coefficients were calculated to explore the associations between the item parameters on the one hand, including item difficulty and item discrimination, and the scored task features on the other hand. A complete overview of all correlation coefficients can be found in Annex 3. Table 11 presents the significant correlation coefficients. For three groups of task features, no significant correlations were found: proposition, inferences, and meta task.

Table 11

Overview of significant results of correlation analyses between item parameters and task features (N = 182)

Features	**Item difficulty (p-value)**		**Item discrimination (rir-value)**	
	r	*p*	*r*	*p*
Type of relationships				
If-then	-	-	.151	.042
Information task				
Matching a point of view with a supporting argument	-	-	-.165	.026
Meta task	-	-	-.197	.008
Drawing a conclusion based on a combination of information elements	-.182	.014	-	-
Implicit language task				
Derive meaning of punctuation marks	-.204	.006	-	-
Explicit language task				
Meaning of words	-	-	.171	.021
Presentation features				
Number of plausible response options (plausible distractors)	-.510	<.001	-	-

First, the results of significant correlations with respect to item difficulty are described, and after, significant correlations with respect to item discrimination. Item difficulty. The (integrative) information task feature *drawing a conclusion based on a combination of information elements* showed a weakly negative correlation with item difficulty ($r(180) = -.182$, $p = .014$). The implicit

language task feature of *deriving the meaning of punctuation marks* is weakly negatively correlated to item difficulty ($r(180) = -.204$, $p = .006$). The presentation task feature, the *degree to which the key is the best possible response* is moderately positively correlated to item difficulty ($r(180) = .350$, $p < .001$).

Item discrimination. Regarding correlation with item discrimination, the latter expressed through the corrected item-total correlation (rir-value), the following task features showed significant correlation coefficients. The (integrative) information task feature *meta task* showed a weakly negative correlation with the rir-value ($r(180) = -.197$, $p = .008$). Another weakly negative correlation with the rir-value was found for the (supporting) information task of *matching a point of view with a supporting argument* ($r(180) = -.165$, $p = .026$). A weakly positive correlation was found for the explicit language task feature of *deriving the meaning of words* ($r(180) = .171$, $p = .021$). Another weakly positive correlation was found for the necessary information relationship feature of the type *if-then* ($r(180) = .151$, $p = .042$).

Table 12

Overview of significant results of analyses of variance, with item parameters as dependent variable

Features	Item difficulty (p-value)				Item discrimination (rir-value)			
	η^2	*F*	df1;df2	Sig.	η^2	*F*	df1;df2	Sig.
Presentation features								
Key is the best possible response	**.122**	**8.266**	**3; 178**	**<.001**	**.124**	**8.386**	**3; 178**	**<.001**
Eliminating response options is necessary	.037	2.252	3; 178	.084	.019	1.144	3; 178	.333
Target skill								
Cluster	.013	0.594	4; 174	.668	**.118**	**5.828**	**4; 174**	**<.001**
Text content								
Name of the corresponding text	**.227**	**2.838**	**17; 164**	**<.001**	**.265**	**3.480**	**17; 164**	**<.001**
Text domain Dutch vocational education	**.057**	**2.658**	**4; 177**	**.034**	**.111**	**5.517**	**4; 177**	**<.001**

Note. Significant ($\alpha < .05$) effect sizes, η^2, are printed in bold print

First, the results of the analyses of variance from Table 12 with respect to item difficulty are reported. Item difficulty. The presentation task feature, the *degree to which the key is the best possible response* showed a medium effect on item difficulty ($F(3,178) = 8.266$, $p < .001$, $\eta^2 = .122$). Another presentation feature, the *necessity for eliminating response options*, did not show a significant relation with item difficulty ($F(3,178) = 2.252$, $p = .084$, $\eta^2 = .037$). Another feature in the group of target skill, is cluster, which also did not have a significant effect on item difficulty ($F(4,174) = 0.594$, $p = .668$, $\eta^2 = .013$). Lastly, in the group of text content task features, both *name of the*

corresponding text and the *text domain* of Dutch vocational education a text belongs to, showed a significant relation with item difficulty. On the one hand, for the *text used*, this magnitude of the relation with item difficulty was large (F (17,164) = 2.838, p < .001, η^2 = .227). On the other hand, for the *text domain*, this magnitude of the relation was small (F (4;177) = 2.658, p = .034, η^2 = .057).

Item discrimination. Regarding the magnitude of the relations with item discrimination, for the following task features significant relations were found. Similarly, as for the presentation task feature the *degree to which the key is the best possible response* and item difficulty, a significant relationship of medium magnitude was found with item discrimination (F (3,178) = 8.386, p <.001, η^2 = .124). Another presentation feature, the *necessity for eliminating response options*, did (similar to item difficulty) not show a significant relation with item difficulty (F (3,178) = 1.144, p = .333, η^2 = .019). Regarding the target skill feature *cluster*, a significant medium relationship with item discrimination was found (F (4,174) = 5.828, p < .001, η^2 = .118).

Finally, in the group of text content task features, both *name of the corresponding text* and the *text domain* of Dutch vocational education a text belongs to, showed a significant relation with item difficulty. On the one hand, for the *name of the text*, the magnitude of the relation found was large (F (17,164) = 3.480, p < .001, η^2 = .265). On the other hand, for the *text domain*, the magnitude of the relation found was medium (F (4;177) = 5.517, p < .001, η^2 = .111).

Text Predicting Item Parameters

Table 13 provides the results of multiple regression analyses with item difficulty and item discrimination respectively as dependent variables and the text as predictor. With regard to *item difficulty,* text (as a basis for items) comes out as a significant predictor. Altogether, 23 per cent of the variance in difficulty is explained by texts (R^2 = .227, R^2_{adj} = .147; F (17.164) = 2.838, p < .001). Especially, items yielded from text N (β = .454; t (164) = 4.778; p = <.001), text K (β = .369; t (164) = 3.602; p = <.001), text D (β = .343; t (164) = 3.431; p = <.001), and text I (β = .323; t (164) = 3.315; p = .001) clearly contain easier items than the other texts. With regard to *item discrimination*, text (as a basis for items) results as a significant predictor of item discrimination. More specifically, some texts show a significant negative contribution to item discrimination, therefore negatively impacting the evidence strength of items subsumed under the text, in order of effect: text E (β = -.320; t (164) = -3.204; p = .002), text C (β = -.235; t (164) = -2.296; p = .023), text J (β = -.216; t (164) = -2.109; p = .036), text B (β = -.202; t (164) = -1.975; p = .050).

Table 13

Results of multiple regression analyses using text, with item parameters as dependent variable

	Item difficulty (p-value)					Item discrimination (rir-value)				
Text	***B***	**s.e.**	**β**	***t*(164)**	**Sig.**	***B***	**s.e.**	**β**	***t*(164)**	**Sig.**
(Intercept)	**45.250**	**5.919**		**7.644**	**<.001**	**19.000**	**2.418**		**7.857**	**<.001**
Text A	**17.750**	**7.642**	**.244**	**2.323**	**.021**	4.583	3.122	.150	1.468	.144
Text B	3.917	7.642	.054	0.513	.609	**-6.167**	**3.122**	**-.202**	**-1.975**	**.050**
Text C	12.000	7.642	.165	1.570	.118	**-7.167**	**3.122**	**-.235**	**-2.296**	**.023**
Text D	**27.250**	**7.942**	.343	**3.431**	**<.001**	-1.400	3.244	-.042	-0.432	.667
Text E	12.386	7.780	.163	1.592	.113	**-10.182**	**3.178**	**-.320**	**-3.204**	**.002**
Text F	**17.750**	**8.371**	**.201**	**2.120**	**.035**	-1.375	3.420	-.037	-0.402	.688
Text G	**17.659**	**7.780**	**.233**	**2.270**	**.025**	-5.182	3.178	-.163	-1.630	.105
Text H	**21.750**	**7.942**	**.274**	**2.739**	**.007**	-2.000	3.244	-.060	-0.616	.538
Text I	**26.972**	**8.135**	**.323**	**3.315**	**.001**	-0.111	3.324	-.003	-0.033	.973
Text J	**15.667**	**7.642**	**.215**	**2.050**	**.042**	**-6.583**	**3.122**	**-.216**	**-2.109**	**.036**
Text K	**28.023**	**7.780**	**.369**	**3.602**	**<.001**	-3.000	3.178	-.094	-0.944	.347
Text L	**22.417**	**7.642**	**.308**	**2.933**	**.004**	2.750	3.122	.090	0.881	.380
Text M	13.750	8.371	.156	1.643	.102	-1.375	3.420	-.037	-0.402	.688
Text N	**40.000**	**8.371**	**.454**	**4.778**	**<.001**	1.875	3.420	.051	0.548	.584
Text O	14.875	8.371	.169	1.777	.077	1.125	3.420	.030	0.329	.743
Text P	11.306	8.135	.136	1.390	.167	-5.333	3.324	-.153	-1.605	.110
Text Q*	-	-	-	-	-	-	-	-	-	-
Text R	**21.659**	**7.780**	**.285**	**2.784**	**.006**	-1.364	3.178	-.043	-0.429	.668

***Text Q is the reference text

Note. Dependent variable item difficulty: $R^2 = .227$, $R^2_{adj} = .147$; $F\ (17.164) = 2.838$, $p < .001$

Note. Dependent variable item discrimination: $R^2 = .265$, $R^2_{adj} = .189$; $F\ (17.164) = 3.480$, $p < .001$

In the sections below, additional task features will be included in regression analyses, in order to predict the item difficulty and item discrimination parameters. These additional task features regarding the specific content of texts will help to explain the differences between texts.

Prediction of Item Difficulty Using All Task Features

In the next sections, the whole set of task features as described in the methods section is systematically used as predictors in multiple regression analyses. The results of the multiple regression analyses are shown in Table 14 and 15. These analyses were conducted using the item sample of 182. In a first analysis, groups of task features as defined in Table 5 are used as predictors. Separate regression analyses were performed with one group of task features as a predictor. Depending on their contribution in the separate analyses, a combined selection of task features was used for subsequent multiple regression analysis.

Table 14

Results of multiple regression analyses using groups of task features, dependent variable: item difficulty

	R^2	R^2_{adj}	*F*	df1; df2	Sig.
Item difficulty					
Proposition	.010	-.012	0.460	4; 177	.765
Type of relationships	.112	-.004	0.962	21; 160	.513
Information task	.080	.031	1.653	9; 172	.104
Types of inferences	.022	.000	1.005	4; 177	.406
Number of inferences	.016	.010	2.889	1; 180	.091
Meta task	.013	-.003	0.803	3; 178	.494
Implicit language task	**.045**	**.029**	**2.822**	**3; 178**	**.040**
Explicit language task	.002	-.001	0.137	2; 179	.872
Presentation features	**.372**	**.319**	**7.067**	**14; 167**	**<.001**
Target skill	.013	-.009	.594	4; 174	.668
Text content – Text	**.227**	**.147**	**2.838**	**17; 164**	**<.001**
Text content – Domain	**.057**	**.035**	**2.658**	**4; 177**	**.034**

The groups of task features that contributed significantly to item difficulty, are shown in Table 14. They include implicit language task (R^2 = .045, R^2_{adj} = .029; *F* (3.178) = 2.822, *p* = .040), item presentation features (R^2 = .372, R^2_{adj} = .319; *F* (14.167) = 7.067, *p* < .001), text content – text (R^2 = .227, R^2_{adj} = .147; *F* (17.164) = 7.606, *p* < .001) and text content – domain (R^2 = .057, R^2_{adj} = .035; *F* (4.177) = 2.658, *p* = .034). The group of presentation features was able to explain the most significant part of variance in item difficulty.

In addition, significant predictors from the analyses on groups of task features were used in a subsequent multiple regression analysis as independent variables. In this analysis, text as a predictor was disregarded. Text was already established as a significant predictor, and the additional task features regarding the specific content of texts will help to explain the differences between texts. The results of this multiple regression analysis can be found in Table 15. Altogether, the task features predicted 44 per cent of the variance in item difficulty (R^2 = .441, R^2_{adj} = .383; *F* (17.164) = 7.606, *p* < .001).

Table 15

Results of multiple regression analyses using significant predictors (excluding text), dependent variable: item difficulty

Task feature	*B*	s.e.	β	*t*(164)	Sig.
(Intercept)	**75.658**	**6.192**		**12.218**	**<.001**
Type of relationships					
If-then	1.782	1.282	.086	1.390	.166
Information task					
Finding justification	**-6.235**	**2.721**	**-.152**	**-2.291**	**.023**
Drawing a conclusion based on a combination of information elements	-5.505	2.843	-.132	-1.936	.055
Separating relevant information	**5.893**	**2.805**	**.153**	**2.101**	**.037**
Implicit language task					
Derive meaning of punctuation marks	**-15.564**	**5.780**	**-.166**	**-2.693**	**.008**
Presentation features					
Key is the best possible response (score = 0)*	-	-	-	-	-
Key is the best possible response (score = 50)	**9.995**	**4.502**	**.254**	**2.220**	**.028**
Key is the best possible response (score = 100)	**13.857**	**4.241**	**.376**	**3.268**	**.001**
Key is the best possible response (NA)	12.044	8.728	.098	1.380	.169
Number of plausible response options (N = 0)*	-	-	-	-	-
Number of plausible response options (N = 1)	**-15.505**	**3.003**	**-.419**	**-5.163**	**<.001**
Number of plausible response options (N = 2)	**-20.985**	**3.418**	**-.467**	**-6.140**	**<.001**
Number of plausible response options (N = 3)	**-42.657**	**8.758**	**-.300**	**-4.871**	**<.001**
Number of plausible response options (NA)	**-10.662**	**3.733**	**-.203**	**-2.856**	**.005**
There exists overlap with other response options	-3.918	2.261	-.108	-1.733	.085
Text content					
Text domain 1: political / legal*	-	-	-	-	-
Text domain 2: economy	-7.910	4.488	-.124	-1.762	.080
Text domain 3: social / societal	-4.378	3.094	-.119	-1.415	.159
Text domain 4: vital citizenship	-5.168	3.658	-.110	-1.413	.160
Text domain 5: career	**-10.152**	**4.037**	**-.190**	**-2.515**	**.013**

Note. R^2 = .441, R^2_{adj} = .383, F (17.164) = 7.606, p < .001

*These factors are treated by R as reference groups

The most powerful predictive task features are presentation features. A presentation feature predicting a higher value for item difficulty (i.e., an easier item) is the presentation feature representing to which *degree the key is the best possible response* (score = 50 and score = 100), respectively β = .254; *t* (164) = 2.220; *p* = .028 and β = .376; *t* (164) = 3.268; *p* = .001. So, compared to the reference group with score = 0, so where the key is coded as not being the best possible response, items for which the key is (partially) the best possible response, are relatively easier. Another item presentation feature, the *number of plausible response options* (plausible distractors) however, predicts a lower value for item difficulty (i.e., a more difficult item). For the number of plausible response options, the reference group was no plausible response options (N = 0). The presence of a number of plausible response options, leads to a lower value for item difficulty (N = 1,

β = -.419; *t* (164) = -5.163; *p* < .001, N = 2, β = -.467; *t* (164) = -6.140; *p* < .001 and N = 3, β = -.300; *t* (164) = -4.871; *p* < .001). The largest beta (β = -.467) was found for two plausible answer options.

Other significant predictors found mostly decreased the p-value, which means making an item more difficult. Firstly, the information task feature *finding justification* (β = -.152; *t* (164) = -2.291; *p* = .023). So, when an item requires a reader to look for a justification, the item difficulty increases. Secondly, when readers, implicitly, had to *derive meaning of punctuation marks,* which is considered a prerequisite language task, in order to correctly answer an item, the item difficulty also increases (β = -.166; *t* (164) = -2.693; *p* = .008). Thirdly, for the text content feature *domain*, one of the domains also predicts an increased in item difficulty. Text domain 5, with a focus on career, causes a decrease in the p-value (β = -.190; *t* (164) = -2.515; *p* = .013), i.e., more difficult items, compared to items from the domain politics and legislation.

A final significant feature is another information task feature. This information task feature, *separating relevant information*, leads to a higher value for item difficulty (β = .153; *t* (164) = 2.101; *p* = .037), i.e., creating an easier item. So, when an item requires a reader to separate relevant information, this apparently found relatively easy.

Prediction of Item Discrimination

The results of the multiple regression analyses with item discrimination as a dependent variable are shown in Table 16 and 17. The procedure of analysing the data was similar to the analyses performed for predicting item difficulty. To start, multiple regression analyses were performed using task features from individual groups as predictors first as a basis of a subsequent analysis with features that had a significant contribution in the separate analyses.

Table 16

Results of separate multiple regression analyses using groups of task features as predictors, with item discrimination as a dependent variable

	R^2	R^2_{adj}	*F*	df1; df2	Sig.
Item discrimination					
Proposition	.021	-.002	0.929	4; 177	.449
Type of relationships	.128	.014	1.119	21; 160	.333
Information task	**.099**	**.051**	**2.091**	**9; 172**	**.033**
Types of inferences	.040	.019	1.853	4; 177	.121
Number of inferences	.000	-.005	.037	1; 180	.847
Meta task	.033	.017	2.027	3; 178	.112
Implicit language task	.011	-.005	0.680	3; 178	.565
Explicit language task	.029	.018	2.696	2; 179	.070
Presentation features	**.165**	**.095**	**2.362**	**14; 167**	**.005**
Target skill	**.118**	**.098**	**5.828**	**4; 174**	**<.001**
Text content – Text	**.265**	**.189**	**3.480**	**17; 164**	**<.001**
Text content – Domain	**.111**	**.091**	**5.517**	**4; 177**	**<.001**

The following groups of task features showed significant contribution to item discrimination, see Table 16: information task (R^2 = .099, R^2_{adj} = .051; *F* (9.172) = 2.091, *p* = .033), presentation features (R^2 = .165, R^2_{adj} = .095; *F* (14.167) = 2.362, *p* = .005), target skill – cluster (R^2 = .118, R^2_{adj} = .098; *F* (4.174) = 5.828, *p* < .001), text content – text (R^2 = .265, R^2_{adj} = .189; *F* (17.164) = 3.480, *p* < .001) and text content – text domain (R^2 = .111, R^2_{adj} = .091; *F* (4.177) = 5.517, *p* < .001). The group text explained the largest amount of variance in item.

In a subsequent regression analysis, all significant predictors that resulted from the separate analyses were used in a combined multiple regression analysis. The results of this analysis can be found in Table 17. Altogether, the task features predicted 43 per cent of the variance in item discrimination (R^2 = .428, R^2_{adj} = .350; *F* (24.154) = 4.998, *p* < .001).

Table 17

Results of combined multiple regression analyses, with item discrimination as a dependent variable

Task feature	*B*	s.e.	β	*t*(154)	Sig.
(Intercept)	**13.191**	**2.667**		**4.946**	**<.001**
Type of relationships					
Effect	**-0.614**	**0.195**	**-.208**	**-3.150**	**.002**
Causal	-0.053	0.540	.007	-0.098	.922
Means-goal	0.651	0.434	.111	1.499	.136
If-then	0.394	0.580	.046	0.680	.497
Others	1.310	0.757	.120	1.730	.086
Information task					
Meta task	-1.715	1.367	-.112	-1.254	.212
Matching a point of view with a supporting argument	-2.301	1.265	-.123	-1.819	.071
Meta task					
Text structure	0.818	1.852	.046	0.442	.659
Explicit language task					
Meaning of words	2.918	2.197	.090	1.328	.186
Presentation features					
Key is the best possible response (N = 0)*	-	-	-	-	-
Key is the best possible response (score = 50)	**6.260**	**1.941**	**.383**	**3.225**	**.002**
Key is the best possible response (score = 100)	**8.377**	**1.811**	**.549**	**4.626**	**<.001**
Key is the best possible response (NA)	**11.330**	**3.836**	**.224**	**2.954**	**.004**
Number of plausible response options (N = 0)*	-	-	-	-	-
Number of plausible response options (N = 1)	-2.060	1.245	-.135	-1.655	.100
Number of plausible response options (N = 2)	-0.964	1.481	-.051	-0.651	.516
Number of plausible response options (N = 3)	-0.707	3.764	-.012	-0.188	.851
Number of plausible response options (NA)	-0.188	1.650	-.009	-0.114	.909
Target skill					
Cluster 1: goal / type of text*	-	-	-	-	-
Cluster 2: text structure	1.822	1.977	.103	0.922	.358
Cluster 3: main and side issues / summarising	0.394	1.850	.019	0.213	.832
Cluster 4: understanding and interpretation of information	**5.050**	**1.721**	**.291**	**2.935**	**.004**
Cluster 5: argumentation	1.319	1.655	.075	0.797	.427
Text content					
Text domain 1: political / legal*	-	-	-	-	-
Text domain 2: economy	-2.298	1.950	-.088	-1.178	.241
Text domain 3: social / societal	**-3.339**	**1.360**	**-.220**	**-2.454**	**.015**
Text domain 4: Vital citizenship	**-5.489**	**1.606**	**-.282**	**-3.417**	**.001**
Text domain 5: Career	**-7.212**	**1.780**	**-.329**	**-4.008**	**<.001**

Note. R^2 = .428, R^2_{adj} = .350; F (24.154) = 4.998, p < .001

*These factors are treated by R as reference groups

The most powerful predictive task features are, again, presentation features. The presentation feature representing to which *degree the key is the best possible response* (score = 50 and score = 100), respectively β = .383; t (154) = 3.225; p = .002 and β = .549; t (154) = 4.626; p < .001. So, compared to the reference group with score = 0, so where the key is not the best possible response, items for which the key is (partially) the best possible response, tend to have increased rir-

values, and therefore have a higher evidence strength. For another task feature, *target skill*, one of the clusters also is a significant predictor for item discrimination. This group specifically is about understanding information and the interpretation of information. Items in this cluster have an increased rir-value, and therefore have a higher evidence strength (β = .291; *t* (154) = 2.935; *p* = .004).

Other task features found as significant predictors for item discrimination predicting a decrease, i.e., creating an item with less evidence strength. Regarding text content features, several *text domains* were found to predict a significantly decrease item discrimination. These were the text domains 3, 4 and 5 regarding social / societal, vital citizenship and career subjects, respectively β = -.220; *t* (154) = -2.454; *p* = .015, β = -.282; *t* (154) = -3.417; *p* = .001, and β = -.329; *t* (154) = -4.008; *p* < .001. All text domains mentioned predict to a decrease in item discrimination, and therefore a decrease in evidence strength of the items in these domains. Finally, a feature from the group of relationship features was also a significant, but the least powerful predictor for item discrimination. This was the relationship *effect* (β = -.208; *t* (154) = -3.150; *p* = .002). This type of relationship has a negative relationship with item discrimination, i.e., an increase of the presence of effect relationships in propositions needed to correctly answer an item, leads to a decrease in evidence strength.

Conclusion and Discussion

Dutch language reading comprehension exams have been topic of investigation recently. The present research focused on finding predictive task features for item difficulty and item discrimination in 3F Dutch comprehension exams. Regarding the task features, twelve groups of task features were distinguished, that fitted into larger categories: features of necessary information (proposition, type of relationships), information task features, including required mental processes, types and number of inferences required, meta task, prerequisite language task, either implicitly required or explicitly asked for, item presentation features (including plausible response option and correctness of the key), intended target skill, text features, including text topic and text domain. In the end, finding task features that might predict item parameters would be beneficial for several reasons. The main benefit, however, is that the construct validity could be affected positively, thereby increasing the coverage of the target skill (reading comprehension), implying a more valid measurement.

This section will first focus on the (groups of) task features found to predict item parameters (item difficulty, item discrimination) in Dutch language reading comprehension exams in previous research. Subsequently, the section will focus on summarising the results of the present study and accounting for them, to clarify the relations between task features and item features. More specifically, this section aims to address the research question focused on the extent to which (groups of) task features can be used to predict item features in 3F Dutch language reading comprehension exams. After discussing the conclusions of the present study, some limitations of the present study will be addressed. Finally, some suggestions for further research are mentioned. The present study is the first of its kind to include the concept map, representing a mental model, in research into reading comprehension.

Predicting Item Difficulty

For predicting item difficulty, previous research uncovered several task features. With regard to the text features, the *type token ratio* and the *number of centralised elements*, were found to increase item difficulty. Other text features, the *total number of elements*, and the *presence of the actor relationship type*, decreased item difficulty. The present study did not find these text features as predictors for item difficulty. The absence of the type token ratio, number of centralised elements, and the total number of elements could have to do with the fact the present research utilised a different method for counting these. The propositional complexity was taken into account on an item level, instead of on a text level.

Regarding the intrinsic task features, a main feature found to increase item difficulty was the *number of inferences to form*. This was not found in the present study, perhaps also due to the fact this measure was taken into account on an item level, as compared to the text level. Other intrinsic task features found to increase item difficulty by previous research were whether an item focused on (1) *assessment of propositions or arguments following the text*, (2) the *amount of necessary information needed to answer an item concerning the entire text, rather than individual passages*, (3), *organisation of information (elements) is central*, (4) *determining rations between numbers,* and (5) *drawing a conclusion based on a combination of informative elements*. One intrinsic task feature found caused a decrease in item difficulty: when an item focused on *retrieving explicit information*. The present study found an association for the following intrinsic task features: whether an item focused on *drawing a conclusion based on a combination of information elements*, further confirming this finding from previous research. However, this feature was not a significant predictor for item difficulty. Another intrinsic task feature did increase item difficulty: when an item focused on *finding justification*. A final intrinsic task feature, which caused an item to become easier, was when an item focused on *separating relevant information*. Apparently, trying to find justification imposes a cognitive load for test takers, causing items to become more difficult. The separation of relevant information, however, causes items to become easier. Perhaps handling information, and choosing the most relevant parts, is easier as compared to combining information elements or finding information elements to support a certain point of view. The other previously found intrinsic task features were not included in the present research, as they did not have enough prevalence in the item sample, so no statements can be made as to their predictive value for item difficulty.

Looking at the access skills, previous research found an increase in item difficulty when *answering an item required an additional (access) skills*, or when test takers had to *deduce meaning from punctuation marks*. The present research also found an association between *deducing meaning from punctuation marks* and item difficulty, more specifically, an item became easier. Whether items *required an additional (access) skill*, was not taken into account in the present study, due to a lack of prevalence in the current item sample.

Considering the item presentation features, previous research found an increase in item difficulty when the *number of substantively plausible distractors* increased. For the presentation feature the *degree to which the key is the best possible answer response*, it was found items became easier. The present study found an association between item difficulty and the *number of plausible response options*, and more specifically, this item presentation feature caused items to become more difficult. For the *degree to which the key is the best possible answer response,* it was found items became easier. The results from the present study are therefore in line with previous research.

Finally, the task features found to impact item difficulty to the largest extent were the item presentation features. The *number of substantively plausible* and the *degree to which the key is the best possible answer response* are eminently a consequence of the item format multiple-choice. The present study might be another indicator that this item format indeed is detrimental for making valid claims about one's reading comprehension skills.

Predicting Item Discrimination

For predicting item discrimination, previous research uncovered several task features. Regarding the text features, it was found that the *number of the contrast relationship type* positively impacted item discrimination. However, the *degree to which relationships are implicit*, and the *number of the non-directly observable characteristics element*, contributed negatively to item discrimination. The present research mostly found task features related to the type of relationship. An association was found for the *presence of the if-then type of relationship*, but this feature was not a significant predictor. The presence of the *effect type of relationship* was a significant predictor, decreasing the item evidence strength.

Considering the intrinsic task features, previous researchers found mostly features that negatively impacted evidence strength: when an item focused on *retrieving social-communicative meaning*, on *matching a supporting argument with a point of view*, and when an item did *not contain a direct information task, but a text transcending task*. On feature found to increase evidence strength was when an item focuses on a *follow-up (or application) task based on textual information*. These findings seem in line with the taxonomy of Bloom, indicating from lower to higher order thinking skills: remembering, understanding, applying, analysing, evaluating, and creating. It seems items concerning higher order thinking skills, such as matching a supportive argument (evaluating), or a text transcending task (creating), lead to less evidence strength. A follow-up, or application task based on textual information, focused on a lower order thinking skill, applying, lead to increased evidence strength. The present study found an association for items focused on matching a point of view with a supporting argument, and items focused on meta task. However, no significant predictors were yielded after analyses. Previous research was concentrated mostly on 2F Dutch comprehension exams, which included an item sample with a good range of rir-values. However, the present research focused on reference level 3F, included a rather small range of (low) rir-values, which could have caused the previous found results to not show.

With regard to the access skills, a task feature found before with a negative impact on evidence strength were items *relating text passages and texts*. Other task features had a positive impact on evidence strength: when the *meaning of words is explicitly asked for*, and when test takers

had to *relate between textual information and common knowledge*. These results might indicate that items focusing on processing larger amounts of information (relating texts passages and texts), thereby increasing test takers' cognitive load, cause a decrease in evidence strength. Items focusing on processing smaller amounts of information (less cognitive load), such as the explicit meaning of words, or the relation between textual information and common knowledge, are related to an increased evidence strength. While the present research found an association between the explicit language task *meaning of words*, and item discrimination, the task feature was not a significant predictor for item discrimination.

When considering target skill features, it was found before that *drawing a conclusion based on (parts of) the text*, increased the evidence strength of an item. The present results were that items focused on the target skill *understanding and interpretation of information* increased evidence strength. It is noteworthy that again, a task feature related to relatively lower order thinking skills, cause an increase in evidence strength.

Looking at the item presentation features, several researchers found the positive impact of the *degree to which the key is the best possible response* on evidence strength. One study also found that when an item *requires eliminating response options*, evidence strength was decreased. Similar to item difficulty, an association with item difficulty and the *degree to which the key is the best possible response* was found in the current study. It was found the higher the *degree to which the key is the best possible response*, the higher the evidence strength of an item. This seems like a logical result, as a more correct key will have more overlap with the response a test taker has formulated in his or her head. The other previous finding, regarding the elimination of response options, could not be replicated in the present study.

Finally, similar to item difficulty, the task features found to impact item discrimination to the largest extent were the item presentation features. The *degree to which the key is the best possible answer response* and when an item *requires eliminating response options* are eminently a consequence of the item format multiple-choice. Considering a similar trend was found for item difficulty, this might be an indicator that the use of multiple-choice items for making valid claims about one's reading comprehension skills is not the best option.

Text (Content) as Predictor for Item Parameters

In the present study, eighteen texts were included. These texts were also included in regression analyses, to check their potential predictive value for the item parameters. It is found that for item difficulty, eleven texts were found to have a significant positive contribution, thereby increasing the p-value, i.e., leading to relatively easier items. Regarding item difficulty, four texts had

a significant negative contribution to item discrimination, thereby decreasing the rir-value, and the evidence strength of the items. This could imply it partly depends on the text whether a constructor can design items that provide strong evidence for reading comprehension. Another implication could be that the variation in task complexity partly depends on the text on which the item is based.

More specifically, the present research also found a relationship for the *text domain Dutch vocational education* with both item parameters. The text domain contributed to items that were slightly easier, and items with less evidence strength. More specifically, the items focusing on the domain of career, were easier compared to the other domains. Items focusing on the domain of either career, vital citizenship, or social / societal, had a decreased evidence strength, compared to other domains.

Limitations of The Present Research and Suggestions for Further Research

Finally, some limitations of the present study need to be mentioned, as well as some suggestions for further research. These are divided by theme: item sample, methods, analyses, and some general suggestions.

Item sample

Firstly, the limitations of the present research concern the item sample of items used to conduct the research. This item sample was rather small. Further research could give more attention to utilising a data set with more items, ensuring a better balance in task features. Also, with regard to the item discrimination, a relatively low rir-value was found, indicating thin evidence. I.e., items do not effectively discriminate between students low and high achieving test takers. In the present research, the somewhat lower rir-value could have a somewhat negative effect on the exploration of task features predictive for the rir-value, as the values were (consistently) low. It is recommended future research focuses on creating an item sample with a broader range of rir-values, also including highly discriminating items (rir-value > .3).

Methods

Secondly, some limitations of the present research have to do with the methods used. The expert concept map was used to operationalise a situation model of the text, and several proposition task features were obtained using these expert concept maps. The construction of the concept map was done by an expert author, and re-checked by two other experts, who also discussed the concepts and relationships as indicated in the concept map with the expert author. Finally, the author of the present research also looked critically and made some adaptations to the concept maps. However, a concept map is still a representation of a person, who is using a certain type of prior knowledge to construct a situation model. It remains a question on whether this concept map is

representative for the situation model constructed by a Dutch vocational education student. By using an expert concept map as a proxy for an expert situation model, the present research attempted to take into account some criticism on the current reading comprehension assessment, i.e., the focus on only certain text passages (Rupp et al., 2006; Snow, 2003). More specifically, this criticism was one of the reasons to look whether items actually cover the full situation model, or only parts of it. Furthermore, using the concept maps to code which propositions were needed to correctly answer an item, the author did this by herself.

Again, there is the limitation of interpretation and prior knowledge, the subjectiveness of the author. It is suggested further research obtains more coders, so inter-rater reliability can be checked. Moreover, in coding which propositions were needed to correctly answer an item, the focus was on the key. However, it is not taken into account that other response options might include distracting elements, therefore, shaking up the situation model for the student. Also, regarding the implicit elements, there was some ambiguity with respect to coding them, as elements might be implicit in one paragraph, but explicit in another. In general, the rule was followed that if an element is mentioned explicitly somewhere in the text, it is considered explicitly. Here, the reasoning was followed that the whole text is necessary to correctly answer an item. However, one could argue that is not the case. Besides, a task feature used in analyses was *other*, from the group of type of relationships. However, *other* encompassed all relationships that could not be classified elsewhere. Therefore, it represents a very diverse group of relationships, and it is hard to interpret the results involving this feature. In addition, it was not taken into account that items are answered sequentially. Therefore, parts of the situation model that were already used in an item before, might be easier to revisit. This is currently not taken into account. To check how test takers interact with the items, a suggestion would be to also include an experimental approach. An example would be to use think-aloud procedures, in which test takers verbalise their thoughts while answering items in different formats

Analyses

Thirdly, some limitations of the present research have to do with the way that the analyses were performed. Most importantly, items were treated as interchangeably, even though they were not (Annex 4). Items are nested within a text, and a multilevel (or test let) analysis might be better suited to account for this “nestedness”. This was also indicated by the fact that some of the texts were significant predictors for the item parameters. This refers back to the first set of limitations of the research, specifically, the number of items in the item sample. There were not enough texts and related items to conduct a meaningful multilevel analysis. Again, for further research, an item sample with more items is recommended.

General suggestions

In addition to the suggestions for further research mentioned in the paragraphs above, some general suggestions for further research are described in this paragraph. The present research focused on information elements, relations among the information elements, and so-called propositions. In this case, only the relationships among the information elements were put into a category (e.g., effect, causal). Further research might focus on the content characteristics of information from a text, to further investigate the possible predictive role of these categories for item parameters. Moreover, the present study did not involve the patterns of interactions between two or multiple features (also, the item sample was too small to obtain reliable results). Future research could focus on discovering these complex patterns of interactions, and their influence on item parameters.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press. https://doi.org/https://doi.org/10.1017/CBO9780511732935

Ausubel, D. P. (1963). Cognitive structure and the facilitation of meaningful verbal learning. *Journal of Teacher Education, 14*(2), 217-222. https://doi.org/https://doi.org/10.1177/002248716301400220

Ausubel, D. P. (2000). Assimilation Theory in Meaningful Learning and Retention Processes. In *The Acquisition and Retention of Knowledge: A Cognitive View*. (pp. 101-145). Springer. https://doi.org/https://doi.org/10.1007/978-94-015-9454-7_5

Cito. (2022a). *Centraal examen mbo Nederlandse taal.* Retrieved October 5 from https://www.cito.nl/onderwijs/middelbaar-beroepsonderwijs/centrale-examens-mbo/centraal-examen-mbo-nederlandse-taal

Cito. (2022b). *Middelbaar beroepsonderwijs.* Retrieved October 5 from https://www.cito.nl/onderwijs/middelbaar-beroepsonderwijs/centrale-examens-mbo

College voor Toetsen en Examens. (2017). *Nederlandse Taal Referentieniveau 3F.* Retrieved from https://www.examenbladmbo.nl/syllabus/syllabus-nederlandse-taal-3f-3/2022-2023/f=/syllabus_Nederlandse_taal_3F_aug_2018.pdf

Deane, P., Sabatini, J., & O'Reilly, T. (2011). *English Language Arts Literacy Framework*. Princeton, NJ: Educational Testing Service. Retrieved from https://www.academia.edu/83775593/English_Language_Arts_Literacy_Framework

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical education, 37*(9), 830-837. https://doi.org/https://doi.org/10.1046/j.1365-2923.2003.01594.x

Goldman, S. R., & Pellegrino, J. W. (2015). Research on learning and instruction: Implications for

curriculum, instruction, and assessment. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 33-41. https://doi.org/https://doi.org/10.1177/2372732215601866

Graesser, A. C. (2015). Deeper learning with advances in discourse science and technology. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 42-50. https://doi.org/https://doi.org/10.1177/2372732215600888

Kendeou, P., Broek, P., Helder, A., & Karlsson, J. (2014). A Cognitive View of Reading Comprehension: Implications for Reading Difficulties. *Learning Disabilities Research & Practice, 29*(1), 10-16. https://doi.org/https://doi.org/10.1111/ldrp.12025

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading Comprehension: Core Components and Processes. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 62-69. https://doi.org/https://doi.org/10.1177/2372732215624707

Kintsch, W., & Rawson, K. A. (2005). Comprehension. In M. J. Snowling & C. E. Hulme (Eds.), *The science of reading: A handbook* (pp. 209-226). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch12

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

Kispal, A. (2008). *Effective teaching of inference skill for reading: Literature review*. National Foundation for Educational Research. Retrieved from https://www.nfer.ac.uk/publications/edr01/edr01.pdf

Lieverse, K. (2021). *Predictive Characteristics of Item-difficulty and discrimination for 2F Dutch Reading Comprehension Exams* [University of Twente].

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012, 13-17 April 2012). *A framework for predicting item difficulty in reading tests.* Annual Meeting of the American Educational

Research Association (AERA), Vancouver. https://research.acer.edu.au/pisa/5/

Messick, S. (1996). Validity of performance assessment. In G. W. E. Philips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). National Center for Education.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i-29.

Mislevy, R. J., & Haertel, G. D. (2007). Implications of evidence-centered design for educational testing. *Educational measurement: issues and practice, 25*(4), 6-20. https://doi.org/https://doi.org/10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *On the Roles of Task Model Variables in Assessment Design.* Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Retrieved from https://cresst.org/wp-content/uploads/TECH500.pdf

Niklas, F., Cohrssen, C., & Tayler, C. (2016). The sooner, the better: Early reading to children. *Sage Open, 6*(4), 1-11. https://doi.org/https://doi.org/10.1177/2158244016672715

Novak, J. D., & Cañas, A. J. (2006). *The Theory Underlying Concept Maps and How to Construct Them* (Technical Report IHMC CmapTools 2006-01, Issue. http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf

Novak, J. D., & Cañas, A. J. (2007). Theoretical origins of concept maps, how to construct them, and uses in education. *Reflecting education, 3*(1), 29-42.

Novak, J. D., & Musonda, D. (1991). A Twelve-Year Longitudinal Study of Science Concept Learning. *American Educational Research Journal, 28*(1), 117-153. https://doi.org/https://doi.org/10.3102/00028312028001117

Ozuru, Y., Rowe, M., O'Reilly, T., & & McNamara, D. S. (2008). Where's the difficulty in standardised reading tests: The passage or the question? *Behavior Research Methods, 40*(4), 1001-1015.

https://doi.org/10.3758/BRM.40.4.1001

Paris, S. G., & Hamilton, E. E. (2014). The development of children's reading comprehension. In *Handbook of research on reading comprehension* (pp. 56-77). Routledge.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading, 18*(1), 22-37. https://doi.org/https://doi.org/10.1080/10888438.2013.827687

Rijksoverheid. (2022). *Referentieniveaus taal en rekenen*. Retrieved October 13, 2022, from https://www.rijksoverheid.nl/onderwerpen/taal-en-rekenen/referentiekader-taal-en-rekenen

Roelofs, E. C., Emons, W. H., & Verschoor, A. J. (2021). Exploring task features that predict psychometric quality of test items: the case for the Dutch driving theory exam. *International Journal of Testing, 21*(2), 80-104. https://doi.org/https://doi.org/10.1080/15305058.2021.1916506

Roelofs, E. C., Keune, K., & Van Hofwegen, L. (2021). *Kenmerken van examenopgaven voor begrijpend Lezen 2F.*

Roelofs, E. C., Postulart, A., Stevens, C., & Keune, K. (2021). *Kenmerken van examenopgaven begrijpend Lezen 3F (mbo) die bijdragen aan itemmoeilijkheid en bewijskracht.*

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441-474. https://doi.org/https://doi.org/10.1191/0265532206lt337oa

Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Rand Corporation.

Snow, C. E. (2003). Assessment of reading comprehension. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 192-218). Guilford.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous and germane cognitive load. *Educational Psychology Review, 22*, 123-138. https://doi.org/10.1007/s10648-010-9128-5

Van den Broek, P., Risden, K., & Husebye-Hartmann, E. (1995). The role of readers' standards for coherence in the generation of inferences during reading. In *Sources of coherence in reading*. (pp. 353-373). Lawrence Erlbaum Associates, Inc.

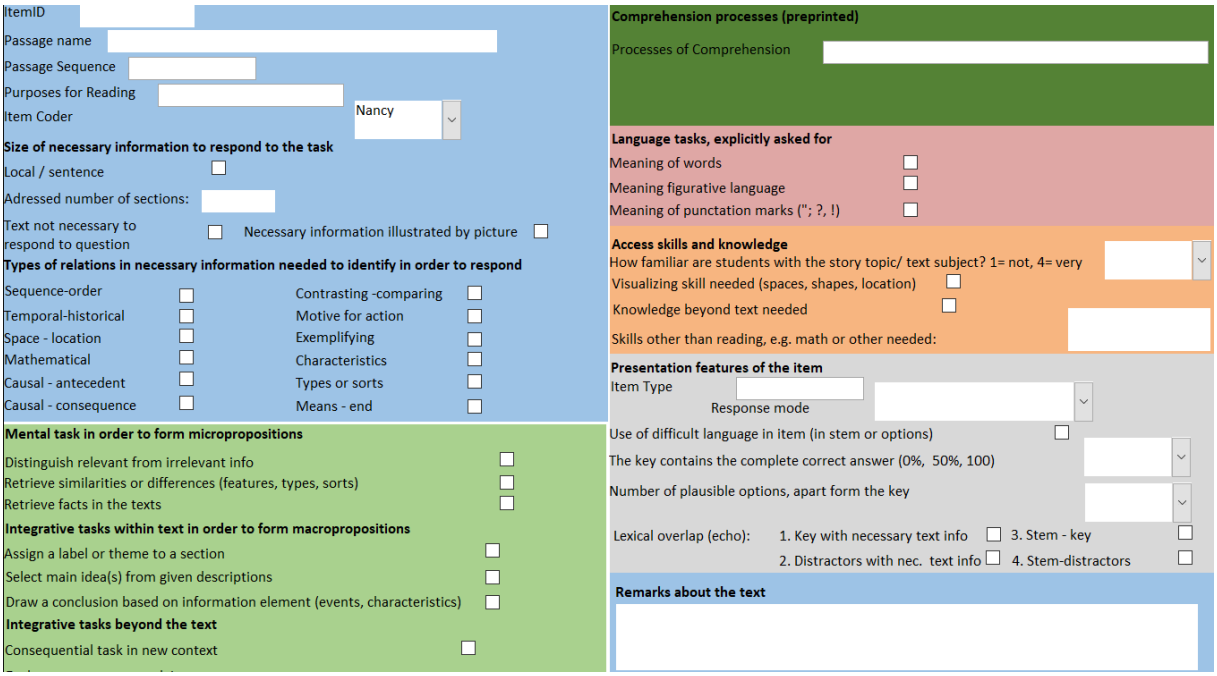
Zijlmans, E. A., Tijmstra, J., van der Ark, L. A., & Sijtsma, K. (2018). Item-score reliability in empirical data sets and its relationship with other item indices. *Educational and Psychological Measurement, 78*(6), 998-1020. https://doi.org/10.1177/0013164417728358

Appendices

Annex 1: Description of the Specified Guidelines for the Reference Level 3F per Skill (College voor Toetsen en Examens, 2017)

Skill	Reference level 3F "The language user is able to …"
Understanding	• Name different types of texts • Express the main idea of a text in his/her own words • Understand and recognise relationships in texts such as cause-effect, means-goal, enumeration, etcetera • Distinguish between essential and side issues, opinions, and facts • Distinguish between point of view and argument • Distinguish between fallacy and argument
Interpreting	• Draw conclusions based on (part of) the text • Draw conclusions about the writer's intentions, views, and feelings
Evaluating	• Can indicate the goal of the author as well as the linguistic means used to achieve this goal • Can divide the text info meaningful units and can name the function of these units • Can assess the acceptability of the argumentation in an argumentative text • Can judge the information in a text on its value to him-/herself and others
Summarising	• Can summarise a text concisely for others
Looking up	• Can assess the reliability of sources and mentions sources • Can quickly find information in longer reports or complicated schedules

Annex 2: Coding Scheme to Determine Task Features per Item in Microsoft Access, Created by Roelofs, Postulart, et al. (2021)



Annex 3: Results of Correlational Analyses and Analyses of Variance Between All Task Features and Item Parameters (N = 182)

	Item difficulty (p-value)	Item discrimination (rir-value)	Item difficulty (p-value)	Item discrimination (rir-value)
	r	*r*	η^2	η^2
Proposition				
Number of propositions needed in order to correctly answer the item	.055	.108	-	-
Number of implicit information elements involved with correctly answering the item	-.014	.052	-	-
Number of implicit relationships between information elements involved with correctly answering the item	.023	.119	-	-
Total number of both implicit information elements and relationships involved with correctly answering the item	-.007	.072	-	-
Number of different types of relationships involved with correctly answering the item	.068	.089	-	-
Type of relationships				
Kind/type	-.098	-.035	-	-
Characteristic/feature	.037	-.008	-	-
Actor	.010	.029	-	-
Effect	-.074	-.143	-	-
Causal	-.059	-.043	-	-
Temporal	-.055	-.050	-	-
Example/illustration	.018	-.013	-	-
Concluding	.139	.068	-	-
Evaluative	-.037	.025	-	-
Means-goal	-.030	.111	-	-
Motive	-.082	-.072	-	-
Explanatory	-.061	-.094	-	-
Location	-.095	-.031	-	-
Source (of information)	.015	-.022	-	-
Recipient	.068	.077	-	-
With respect to	-.042	-.032	-	-
Contrasting	-.033	.030	-	-
Condition	-.041	-.026	-	-
If-then	.116	**.151**	-	-
Implying	.043	.023	-	-
Others	-.113	.107	-	-
Information task				
Finding justification	-.133	-.136	-	-
Similarities and differences	-.066	.066	-	-
Matching a point of view with a supporting argument	-.112	**-.165**	-	-
Separating relevant information	.011	-.044	-	-
Searching for literal information	-.043	-.026	-	-
Meta task	-.024	**-.197**	-	-
Selecting a written summarising sentence	.054	-.039	-	-

Finding arguments (including advantages and disadvantages)	-.053	-.060	-	-
Drawing a conclusion based on a combination of information elements	**-.182**	-.031	-	-
Types of inferences				
Connecting information	-.124	-.081	-	-
Filling an information gap with factual knowledge	-.056	.099	-	-
Deriving new information by logic reasoning	-.014	.106	-	-
Finding a subordinate label	-.113	-.121	-	-
Number of inferences				
Number of inferences	-.126	-.014	-	-
Meta task				
Argumentation theory	-.062	-.066	-	-
Text structure	-.063	-.113	-	-
Social communicative meaning making	.077	-.082	-	-
Implicit language task				
Meaning of words	-.078	-.012	-	-
Expression(s)	-.024	-.043	-	-
Derive meaning of punctuation marks	**-.204**	-.101	-	-
Explicit language task				
Meaning of words	.038	**.171**	-	-
Expression(s)	-.002	.032	-	-
Presentation features				
Key is the best possible response	-	-	**.122**	**.124**
Eliminating response options is necessary	-	-	.037	.019
Number of plausible response options (plausible distractors)	**-.510**	-.103	-	-
Overlap between key and textual information mentioned	.117	.064	-	-
Overlap distractors with textual information mentioned	-.001	.053	-	-
Overlap between item stem and the key	-.045	.037	-	-
Overlap between item stem and distractors	-.116	-.042	-	-
Target skill				
Cluster	-	-	.013	**.118**
Text content				
Name of the corresponding text	-	-	**.227**	**.265**
Text domain Dutch vocational education	-	-	**.057**	**.111**

Note. Significant effect sizes (r/η^2) ($\alpha < .05$) are printed in bold print

Annex 4: Description of the Results of One-way Analyses of Variance (ANOVA)

A one-way ANOVA was conducted to compare the mean p-value between texts. There was a significant difference found between groups (F (17.164) = 2.838, p <.001). Another one-way ANOVA was conducted to compare the mean rir-value between texts. (F (17.164) = 3.480, p <.001).