The development of the Diagnostic Educational Test (2012-2017)

Sanneke Schouwstra (editor)

The development of the Diagnostic Educational Test (2012-2017)

Sanneke Schouwstra (editor)

Translation: Radboud in’to Languages

Cito | Arnhem

Authors: Arjan Aarnink; Paul Drijvers; Joke Hofstee; Laura van Hofwegen; Herbert Hoijtink; Kirsten van Ingen; Karen Keune; Patrick de Klein; Jesse Koops; Roelien Linthorst; Peter van Os; Daniel van der Palm; Sanneke Schouwstra; Uriël Schuurs; Irene van Stiphout; Wilma Vrijs; Ingrid Williams

Editor: Sanneke Schouwstra

Screening: Sandrijn Vernooij

Translation: Radboud in’to Languages

With co-operation of:

Assessment experts

English

Margreet van Aken; Jonna Bos; Rick Godschalk; Evelyn Reichard; Natalie Schols; Ingrid Williams, Wilma Vrijs

Dutch

Ron van den Beemt; Tom Duindam; Hiske Feenstra; Karin Heesters; Laura van Hofwegen; Kirsten van Ingen; Roelien Linthorst; Uriël Schuurs

Mathematics

Saskia van Boven; Nynke Brandsma; Madelon Groenheiden; Kees Lagerwaard; Ger Limpens; Melanie Steentjes; Irene van Stiphout; Paul Drijvers

Support staff

Frederik de Boer; Esmeralda Craamer; Jacqueline Drijber; Monique Esvelt; Jorien van Haaften; Delphine Kerkmeester; Linda Ravelli; Merle Spruijt; Evelien Verhelst; Sandrijn Vernooij; Wietske Vriezen; Rieneke van der Wal

Scientific researchers

Servaas Frissen; Herbert Hoijtink; Karen Keune; Jesse Koops; Daniël van der Palm; Ivailo Partchev; Erik Roelofs, Sanneke Schouwstra

IT

Arjan Aarnink; Floris Briolas; Wilco Budding; Stefan Garcia; Martijn van Haren; Martijn Hendriksen; Marcel Hoekstra; Nard van Kessel; Patrick de Klein; Marcel Lagerwerf; Mark Martinot; Peter van Os

Management and project management DTT

Anne-Marie Anthonissen; Godelieve van den Boogaard; Joke Hofstee; Sanneke Schouwstra; Lody Smeets; Wilma Vrijs; Anke Weekers

Steering group

Anton Beguin; Aukje Bergsma; Alice Groen; Bert IJsveld; Ed Kremers; Paul van der Molen; Mark Molenaar; Cor Sluijter

With thanks to:

Members of the construction groups:

English

Arthur Gelink; Marieke Nijhof; Lotte Poll; Merel Thijink; Aline van Engelenburg; Nicolette Hardon; Hans Hillen; Jacqueline Pantophlet; Hafize Topcu; Maleene Wilson

Dutch

Rinske Groenier; Irma Nelck; Miranda Paulen; Albert Pouwels; Renée van de Schans; Manon Tiehuis

Mathematics

Ans van der Ark; Jeroen Duineveld; Sietske Greeuw; Mark Henkens; Marieke Hoving; Mascha Klerx; Jan-Otto Kranenborg; Eric Lemson; Frits Lemstra; Lianne Maerman; Martin Mollema; Gerrit Ros; Wendy Smit; Lieke Verschueren

Standard-setting experts:

English

Martina Acharrat - Winter; Andrea Barthel; Iris Berends; Wilfred Berkhof; Jos Bohnen; Ioana Bona; Adrienne Buschmann; Bert Christiaans; Rukiye Dutar; Yaira Eekhof; Jiska Evers; Nagad Farah; Gieneke Former; Aukje de Gier; Martien de Graaf; Monique Harleman; Helen Hol; Claudia Huiskes; Loes Janssen; Evelien van Kats; Sven Klippel; Jan Kortman; Joyce Marien; Hetty Mulder; Salina Nowak; Kitty van Ool; Jacqueline Pantophlet; Nancy Pinas; Iris Praster; Anke van den Reek; Mirvat Salman; Edwin van Schie; Tim Simons; Marjolijn Spijker; R. Tieben; Nyree Vermey; Kitty Weterings; Deidre van de Wijdeven; Marriet Wijma

Dutch

L. Aarab; Mieke Balemans; Peter van Ballegooijen; Martin Bouma; Frederique van Breugel; Robert Chamalaun; Claudia Delfgou; Annelies van Diepen; Kelly Diepenbroek; Marije Doedens; Gjalt van Duinen; Ank Gakpo-van Bommel; Esther van Gorkum; Rien de Grauw; Marc Groen; Miranda Hansen; Waldo Happee; Miriam Heijs; Sonja Heynsdijk-Koch; Sarah Hock; Frans Hoyinck; Marian Kleverkamp; Maud Koenen; Mara Kuijpers; Miranda van Loon; Anne-Marie Maartens; Jaap Molenhuis; Sandra Molhoek; Sonja Mombarg; Arva Muradin; Nienke Nagelmaeker; Rachel Ooft; Bas van Pelt; Mariska Prevoo; Sophie Raaijmakers-Rietbroek; Larissa Rutten; Angela Schellekens; W.R. Scheppink; Janneke Sleenhof; Kim van Soest-Meijer; Jos van Son; Liesbeth Stolk; William Verkade; R. Verkerke; K.R. Verschuren; Lydia van Vliet; Inge de Vries; Elly Wilting

Mathematics

Ajrir; Azougagh; Carla Beckschebe; Mohammed Beladel; Berghuis; A. Brinker; Karel Broekhuis; K. van Camfort; Chi Ki Cheung; Coen van Diepen; Martha van Eijk; Nienke Flier; Claudia van Geffen; Madelon Groenheiden; Annie van der Heijden; J.C. Heilig; E. Heitmeijer; Floortje Holten; Jongerius; Soraya Kamps-Soebhag; Edward van Kervel; Fayola Klaris; Gert de Kleuver; Michelle Kuijt; Ton Lecluse; H. van Loon; Marko Meijer; Mariska Mesken; Majelle Meulman; Saro Mottes; Elena Maria Fonseca Neves; Marion van den Nieuwenhuijzen; Marleen Peddemors-Greftenhuis; Wineke Ritzema; H.M.J. Rorink-Heerink; C. J. Smit; Wybe Stavenga; Roel Veerbeek; Jade Vlaming; Jurgen van der Wal; Annemarie Westerbeek; Mark Wiendels; Michael van Woerkom; Vincent Ykema; A. Zomerman; Jan-Willem Zuijderduijn

Table of contents

1 Background and test-development process of the Diagnostic Educational Test (DET)

Sanneke Schouwstra

1.1 Background of the DET

The idea for the development of a diagnostic educational test (DET) emerged from a recommendation issued by the Education Council of the Netherlands in early 2011. The Education Council recommended that schools should determine the status of all students at the end of the second academic year, in order to monitor their level halfway secondary education. “Where necessary, the interim assessment of the level should lead to the adjustment of the educational program for individual students or groups of students” (Education Council, in Dutch *Onderwijsraad*, February 2011, p. 14). This recommendation was followed up in the Action Plan for Better Performance (*Actieplan Beter Presteren,* May 2011). This plan states that the then Minister of Education, Culture, and Science, Marja van Bijsterveldt-Vliegenthart, wished to implement a diagnostic educational test for the subjects relevant to progression (the so-called core subjects), as a stimulus for the further realization of a culture of data driven teaching, which focuses on optimizing the learning results of all students.

On June 19, 2013, the then State Secretary of Education, Culture, and Science, Sander Dekker, submitted a legislative proposal for the implementation of a student tracking system and a diagnostic educational test in secondary education. According to the proposal (2013, p. 9), the implementation of the diagnostic educational test (DET) would be compulsory for all students in all forms of education in the second year of preparatory secondary vocational education (vmbo) and in the third year of senior general secondary education (havo) and pre-university education (vwo)[1]. This DET had a primarily formative function (2013, p. 14). The DET was focused on collecting as much information as possible for teachers with regard to the performance of students, so that teachers could take more targeted action for improvement and customization (p. 23).

On commission from the Ministry of Education, Culture, and Science, Cito began developing the Diagnostic Educational Test in 2012, under the direction of the Board of Tests and Examinations (in Dutch, *College voor Toetsen en Examens*, or CvTE). The DET was to be based on the educational interim objectives and a description of subject-based content and skills in an ongoing learning line (SLO, 2012). Further, it was to be administered in Facet, the digital system for the administration of centralized tests and examinations (managed by the Executive Agency for the Department of Education [DUO]).

Following a lively political debate on the utility and necessity of testing in basic education, the Ministry of Education, Culture, and Science issued a revised assignment to the Board of Tests and Examinations in 2014 for the development and implementation of the Diagnostic Educational Test. According to this assignment, the DET was to be further developed in a three-year pilot project, together with schools, on a voluntary basis. A prototype of the DET was to be delivered at the end of the pilot period.

In June 2016, a general consultation was held between the standing committee for Education, Culture, and Science and the State Secretary, in which it was decided that the prototype DET would be transferred to the market at the end of the pilot project (Parliamentary document: report from the general consultation, August 17, 2016). In the subsequent months, the ministry engaged in dialogue with various providers of tests and teaching materials. For example, information sessions were organized for interested parties, including presentations and demonstrations on the DET. In early June 2017, it was announced that two organizations were willing to offer a Diagnostic Educational Test starting in the 2017-2018 school year. Delivery of the DET was started six months before the end of the pilot period (in June 2017), thus allowing the Ministry of

[1] In the Netherlands are three kinds of secondary education: (1) pre-vocational secondary education (vmbo) which takes four years - ISCED 2; (2) senior general secondary education (havo) which takes five years - ISCED 3; and (3) pre-university education (vwo) which takes six years - ISCED 3.

Education, Culture, and Science to transfer the assessment to market parties. This report describes the development of the DET prototype delivered in 2017.

1.2 Underlying principles and properties of the DET

The underlying principles for the DET are described in the legislative proposal Student-tracking system and the Diagnostic Educational Test (Parliamentary Paper [*Kamerstuk*] 33661, 2013). The proposal called for a compulsory national DET for three subjects relevant to progression: Dutch language, English language, and mathematics, as well as with regard to arithmetic. The test was intended for all students at the end of the lower years of pre-vocational secondary education, called vmbo (second year) and senior general secondary education, called havo, and pre-university education, called vwo, (third year). The method-independent, nationally normed test (legislative proposal, p. 19) was to be provided at the levels of the basic vocational program (vmbo-bb), the middle-management vocational program (vmbo-kb), and the combined and theoretical programs of pre-vocational secondary education (vmbo-gt), the level of senior general secondary education (havo), and the level of pre-university education (vwo)[2].

The DET had a formative function, and it was diagnostic in nature. The educational interim objectives, which the Netherlands Institute for Curriculum Development (SLO) yet had to determine, were to provide direction for the content of the DET. The interim objectives were to be formulated at five levels of mastery, rooted in the core educational objectives of the lower years, the reference levels, the final attainment levels, and, for English, the European Framework of Reference (p. 24).

Another specification was that the possibility should be investigated of administering the test in adaptive and digital form (p. 23, p. 42). The notion of a diagnostic educational test was to be elaborated further and tested for feasibility in a preliminary study. Such a preliminary study was regarded as highly important to the success of the assessment, given the complexity and challenge of the instrument, which would call for a great deal of innovative ability on the part of the developers.

1.2.1 Properties of the DET

Formative function

The instruments used for gathering information in formative evaluation are not necessarily different from “summative” tests. When test results are used to place, select, certify, or classify students, the tests are said to serve a summative function. Tests serve a formative function if their results are used to provide feedback and to provide further guidance and direction to the learning process (Heritage, 2007; Roelofs & Schouwstra, 2012; Van der Kleij, Vermeulen, Eggen, & Veldkamp, 2013).

Several strategies exist for obtaining information for formative evaluation, including spontaneous evaluations during a class and planned assessment points during classes (Heritage, 2007). One characteristic of such strategies for obtaining information is that they are related to the teaching methods used by the teacher. The DET is a different form of formative evaluation: method-independent formative evaluation (Roelofs & Schouwstra, 2012, p. 17).

Diagnostic nature

In order to serve this formative function, the information must be collected using a diagnostic test that charts the learning needs of students. Diagnostic tests were originally intended primarily for the identification of students with special educational needs (Van der Kleij, Vermeulen, Eggen, & Veldkamp, 2013). For example, they could identify students with learning problems or identify outstanding students, who were in need of greater challenge. One property of the DET is that the test is intended for all students. Each student has unique educational needs. Responding to those needs can help each student develop to their fullest potential.

Several types of diagnoses can be distinguished and used as a base for formulating intervention recommendations. First, there are clarifying diagnoses. Does the student have an overall mastery of the

[2] Hereafter, these types of secondary education (vwo, havo and vmbo) and educational programs (vmbo-gt, vmbo-kb and vmbo-bb) will be referred to as educational streams

specified educational objectives or not? This is the type of diagnosis that is usually provided by summative tests. Refined diagnoses (the second type) are used to generate a profile of strengths and weaknesses. Does the student possess the knowledge, sub-skills and strategies needed to perform at a given level? The third type of diagnosis is the explanatory diagnosis, which indicates why the student is strong or weak in a specific area. Such explanatory diagnoses constitute a far-ranging ambition that requires thorough scientific, psychological, and pedagogical research and knowledge. Within the DET, therefore, we focused primarily on the second type of diagnoses: refined diagnoses.

There are various types of diagnostic tests (Rupp, Templin and Henson, 2010). Although many diagnostic tests are bound to specific teaching methods (Oomens et al., 2017), there are also “cognitive diagnostic tests.” These tests are designed to provide more detailed information on the cognitive strengths and weaknesses of students (Leighton & Gierl, 2007). Such tests are based on cognitive models or student models. A student model describes the nature of the focal skill. The DET is such an assessment based on student models. The model describes a cohesive structure of knowledge aspects, sub-skills and/or strategies that are needed for performance at a given level or that are part of the skill.

Adaptive test

In order to provide such detailed information and many diagnoses, it was necessary for the DET to be adaptive. Ordinary (linear) tests require a large number of tasks and student answers in order to allow all of the necessary diagnoses. It is practically infeasible, however, to administer such amounts of tasks to students. Adaptive tests allow more effective testing. An adaptive test is one that automatically adjusts to the response behavior of individual students. After each item that is answered, a determination is made of what is already known about the strengths and weaknesses of the students and which items must still be presented in order to arrive at an accurate diagnosis.

A second reason for making the DET adaptive was that the original idea had been to enable students to demonstrate if they had a higher level of mastery in particular subjects (legislative proposal, p. 23). After the revised assignment in 2014, however, the Ministry of Education, Culture, and Science decided that it would not be necessary to include such “upstream indications”[3] in the prototype, as such indications would require additional administrations (in the second year of havo).

Digital administration

A fully adaptive administration, in which each task is followed by an evaluation of what is known about the student up to that point and which tasks should be presented subsequently, requires digital administration and the automatic evaluation of student answers. The automatic evaluation of student answers also makes it possible to have the test results ready almost immediately after the test has been administered. The digital administration, the automatic evaluation, and the rapid reporting of outcomes should also reduce the burden on teachers (legislative proposal, p. 22).

Digital administration also offers benefits to students, who have grown up in the digital era. Many schools currently work with digital teaching tools, tablets and computers. Digital testing is therefore better suited to the life experiences of students. It also makes it possible to develop new forms of tasks, which make greater use of the digital possibilities to which students are currently accustomed.

In 2011, on commission from the Ministry of Education, Culture, and Science, construction was started on Facet, the software for the administration of digital examinations and tests (College voor Toetsen en Examens, 2017). The intention was for the Diagnostic Educational Test to be administered within this new digital administration system.

[3] Indication that a student can participate in a higher level education, e.g. instead of pre-vocational secondary education (vmbo) can follow senior general secondary education (havo)

Box 1-1. Overview of the underlying principles and properties of the DET

- For three subjects relevant to progression: Dutch language, English language, and mathematics/arithmetic
- At the levels vmbo-bb, vmbo-kb, vmbo-gt, havo and vwo
- For all students at the end of the lower years of vmbo (second year) and havo/vwo (third year)
- Method-independent, nationally normed
- Formative function
- Diagnostic nature:
  based on student models: for all students, a diagnosis of recognition for the knowledge aspects, sub-skills and strategies needed in order to perform at a given level
- Adaptive
- Digital (in Facet)

1.3 The development process of the assessment

1.3.1 Assessment design

The development of the assessment started with a preliminary study elaborating the concept of a diagnostic educational test and assessing the feasibility (Roelofs & Schouwstra, 2012). A test-development approach was established for the development of the DET, corresponding to evidence-centered design (ECD; Mislevy, Steinberg & Almond, 1999; Mislevy, Almond & Lukas, 2003). The ECD framework forces designers to construct clear argumentation for all activities and processes involved in the assessment of complex skills (Roelofs & Schouwstra, 2012).

According to Mislevy, Almond, and Lukas (2003), test administration can be depicted as a cycle of several processes (see Figure 1-1). The assessment design should include the development of a blueprint for each of these processes (see Figure 1-2). During the preliminary study, a draft assessment design was created by assessment experts in the subjects English, Dutch, and mathematics, along with psychometricians, educational researchers, and ICT specialists. Multiple sub-models were distinguished within the assessment design and constructed simultaneously: the student model, the response-processing model, the measurement model, the task model, the reporting model, the assembly model, the presentation model, and the administration model (see Figure 1-2).

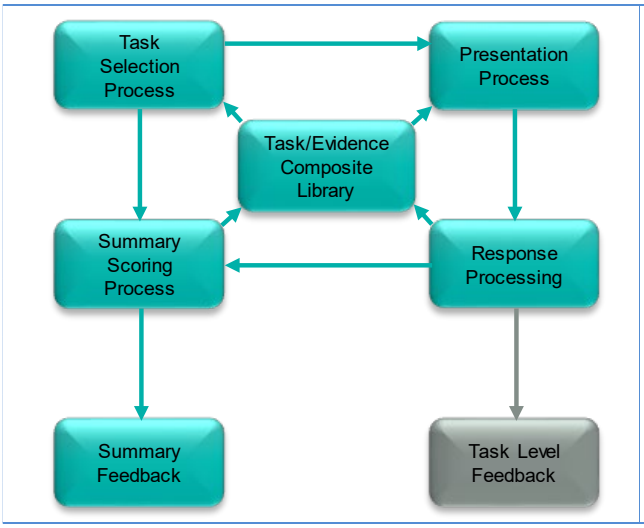


Figure 1-1. Processes in the assessment cycle (adapted from Mislevy, Almond, & Lukas, 2003)

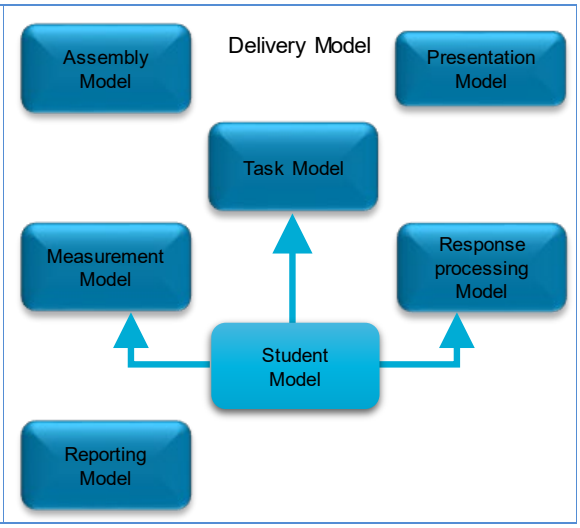


Figure 1-2. Assessment design

- The *student model* constitutes the core of the framework. As stated previously, the student model describes the knowledge, sub-skills, and/or strategies needed in order to perform at a given level.
- The *task model* is the foundation for task development. It describes the types of tasks and situations that can be used to observe the student's skills, as well as the responses that can be expected of students.
- The *presentation model* describes how these tasks are presented to students.
- The *response-processing model* describes how the responses are evaluated and how answers are coded or scored (in the ECD framework called evaluation rules).
- The *measurement model* connects the observations to the relevant skills and describes the standard setting. In the ECD framework of Mislevy, Steinberg and Almond (1999), the response-processing model (scoring) and the measurement model are combined into a single model: the evidence model.
- The *assembly model* describes the selection of tasks and the composition of the assessment. In the context of the Diagnostic Educational Test, the assembly model thus describes the adaptive administration.
- The *reporting model* describes how the report is designed, such that it satisfies the user requirements and the function of the test. Although there is no reporting model within the ECD framework of Mislevy, Steinberg and Almond (1999), reporting is crucial to fulfilling the formative function. It was therefore added to the design of the DET.
- The *delivery model* describes how the various models work during an assessment delivery. It also describes aspects that span all of the other models (e.g., the platform on which the test is administered).

1.3.2 Phases in the test development

The entire process of developing the DET was phased (see Figure 1-3). Subject-area experts, and particularly teachers and schools, were involved throughout the entire process of developing the DET.

1. The first phase involved the formulation of the interim educational objectives by the Netherlands Institute for Curriculum Development (SLO, 2012). For purposes of formulating these educational objectives, SLO consulted subject-area experts (e.g., through a digital survey).
2. This was followed by the preliminary study (Roelofs & Schouwstra, 2012), in which the concept of a diagnostic test and the test-development approach were elaborated.
    a. The student models were developed first. The interim objectives were analyzed in order to identify relevant aspects of knowledge and skills for the student model. Consultations of the educational sector on the first draft of the student model were also conducted. Finally, the definitive student models were elaborated by the Assessment Specification committees of the Board of Tests and Examinations (in Dutch, *College voor Toetsen en Examens*, or CvTE), which consisted of subject-area experts.
    b. A draft assessment design was subsequently developed.
3. Based on the initial assessment design, a try-out was conducted, in which prototypical tasks were developed and tested in small-scale administrations in which schools participated on a voluntary basis. The findings were used to refine the assessment design and to elaborate the specifications. During the try-out, further investigation was conducted on quality assurance (e.g., simulation research to determine the number of tasks needed).
4. The try-out was followed by the large-scale development of tasks with construction groups consisting of subject-area teachers, and the item bank was filled. The confirmation committee of the Board of Tests and Examinations determined which tasks could be tested in the pre-test.

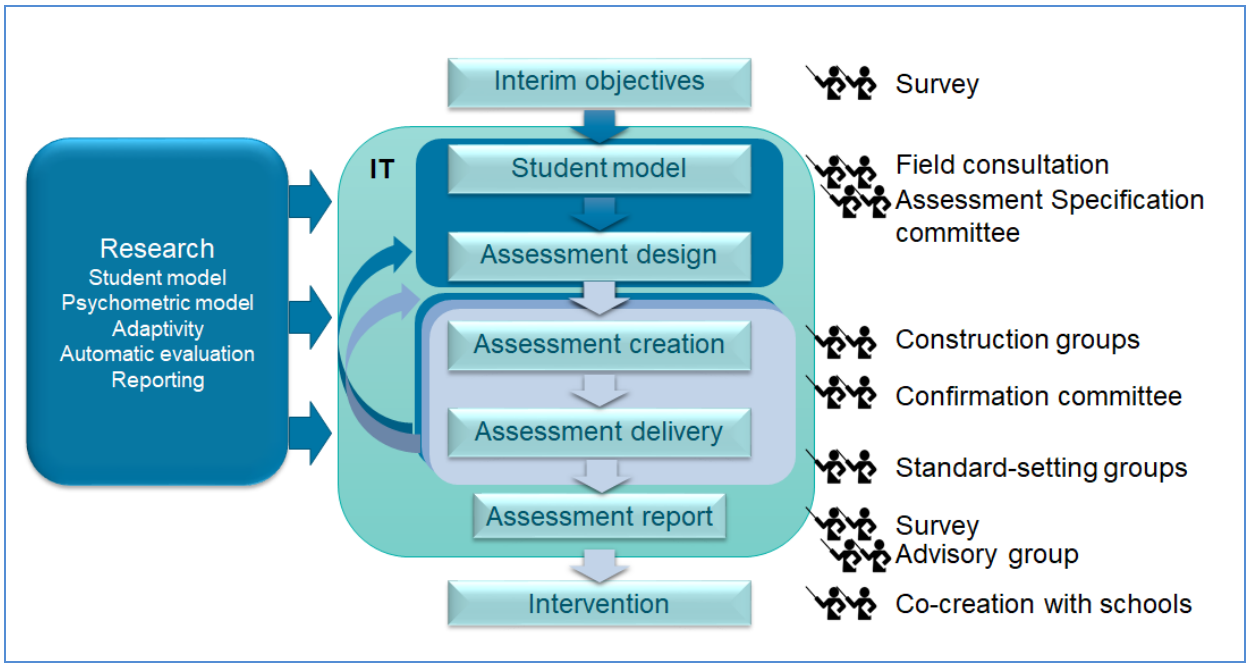


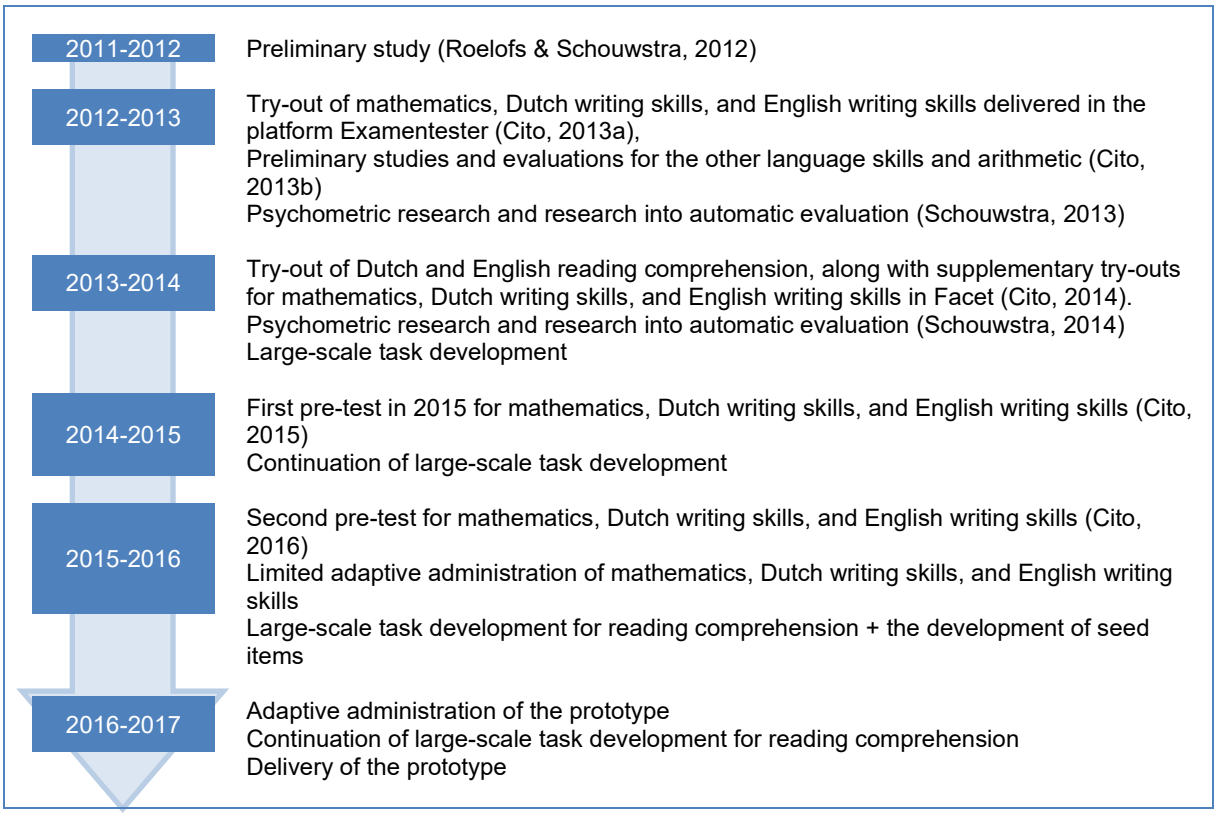
Figure 1-3. Development of the DET

5. The confirmed items were administered in two pre-tests. Thereafter the items that were administered were used in a test-oriented standard setting, in which teachers were involved. The resulting standards and the pre-test data were used to calibrate the tasks.
6. The first administration was then performed in the intended structure and infrastructure.
7. After the first administration the operational phase can follow, in which the test will be administered and maintained each year.

Some phases proceeded in part simultaneously, as the legislative proposal called for the rapid realization of the DET. According to the legislative proposal, the initial administration was to take place in 2015. The plans were paced following the revised assignment of the Ministry in 2014, however, and the initial, limited adaptive administration was held in 2016.

The development of the DET was implemented using a growth model. According to this growth model, a phase was to cover one year, and the development of a new language skill would be started each year. After the revised assignment of the Ministry in 2014, it was decided to pre-test the tasks that had already been developed in two smaller administrations with voluntary participation, instead of in one extensive administration. The further development of the reading-comprehension assessment was also paced, and the development of listening-skills assessment was discontinued. Following the decision to transfer the DET, the Ministry of Education, Culture, and Science decided to discontinue the pre-testing of tasks for reading comprehension, because this would not result in a fully completed part of the prototype. The following figure provides an overview of the phases followed in each school year (Figure 1-4).

Figure 1-4. Timeline for the development of the DET



1.4 Realization of the assessment design

The next chapters describe the elaboration of the assessment design into the prototype that was delivered in 2017. The first part of the report provides an explanation of the approach to task development (Chapter 2) and response processing (Chapter 3) across all three school subjects. This is followed by descriptions of the student models and operationalization (i.e., the realization of the task models) for each subject separately. Special attention is devoted to open-ended items for languages and mathematics.

In Chapters 4 to 8, the development is discussed for each subject separately. For purposes of readability some overlap occurs. In Chapter 4 the development of the DET for Dutch is discussed and in Chapter 5 the development of the DET for English. Chapter 6 discusses the study of the assessment of open-ended writing tasks for Dutch and English. After the languages, Chapter 7 addresses the development of the DET for mathematics, and Chapter 8 focuses on the development of automated assessment for mathematics.

The third part of the report explores technology and psychometrics. First, Chapter 9 describes how information is exchanged between the authoring environment (which contains the tasks) and the digital administration environment, Facet. This is followed by a description of the processing and analysis of the data that were returned (Chapter 10), the standard setting (Chapter 11), and the psychometric model underlying the assessment outcomes (Chapter 12). In Chapter 13, the adaptive assessment delivery is explained.

The last part of the book addresses the results. Chapter 14 describes how a new form of reporting was developed for the assessment outcomes. Finally, the results of the fully adaptive administration in 2017 are presented (Chapter 15). The report closes with some concluding thoughts (Chapter 16).

1.6 References

Cito (2013a). *Diagnostische tussentijdse toets: Verslag try-out 2013 [Diagnostic Educational Test: Try-out report 2013].* Arnhem: Cito.

Cito (2013b). *Diagnostische tussentijdse toets: Voorstudies en evaluaties 2013 [Diagnostic Educational Test: Preliminary studies and evaluations 2013].* Arnhem: Cito.

Cito (2014). *Diagnostische tussentijdse toets: Verslag try-out 2014 [Diagnostic Educational Test: Try-out report 2014].* Arnhem: Cito.

Cito (2015). *Diagnostische tussentijdse toets: Verslag pretest 2015 [Diagnostic Educational Test: Pre-test report 2015].* Arnhem: Cito.

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016 [Diagnostic Educational Test: Pre-test report 2016].* Arnhem: Cito.

College voor Toetsen en Examens [Board of Tests and Examinations] (2017). *Facet geslaagde examensoftware [Facet: Successful examination software].* Utrecht: College voor Toetsen en Examens.

Heritage, M. (2007). Formative Assessment: What Do Teachers Need to Know and Do? *Phi Delta Kappan*, *89*, 140-145.

Leighton, J. & Gierl, M. (Eds., 2007). *Cognitive Diagnostic Assessment for Education: Theory and Applications.* Cambridge University Press.

Mislevy, R. J., Almond, R. G. & Lukas, J. F. (2003). *A Brief Introduction to Evidence-centered Design*. Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-Centered Assessment Design*. Princeton, NJ: Educational Testing Service.

Onderwijsraad [Education Council] (2011). *Advies: Naar hogere leerprestaties in het voortgezet onderwijs. [Advise: Towards higher achievements in secondary education].* Den Haag: de Onderwijsraad.

Oomens, M., Exalto, R., de Jong, A., Scholten, F., Veldkamp, B., Janse, R. & Scheerens, J. (2017). *Marktonderzoek formatief evalueren: Een onderzoek naar vraag en aanbod [Market research on formative evaluation: A study of supply and demand].* Oberon.

Parliamentary document. *Verslag van een algemeen overleg, gehouden op 14 juni 2016, over de diagnostische tussentijdse toets [Report on a general consultation on the Diagnostic Educational Test]* (in Dutch) (2016, 17 14 AugustJune).. Retrieved from https://www.tweedekamer.nl/debat_en_vergadering/commissievergaderingen/details?id=2016a02431

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic Measurement: Theory, Methods, and Applications*. New York: The Guilford Press.

Schouwstra, S. (Ed.) (2013). *Diagnostische tussentijdse toets: Onderzoek 2013 [Diagnostic Educational Test: Research 2013].* Arnhem: Cito.

Schouwstra, S. (Ed.) (2014). *Diagnostische tussentijdse toets: Onderzoek 2014 [Diagnostic Educational Test: Research 2014].* Arnhem: Cito.

SLO: Nationaal expertisecentrum leerplanontwikkeling [Netherlands Institute for Curriculum Development] (April 2012). *Concept-tussendoelen kernvakken onderbouw vo [Draft interim objectives for core subjects in the lower years of secondary education].* Retrieved from http://www.slo.nl/nieuws/00291/.

Van Bijsterveldt-Vliegenthart, M. (2011, May 5th). *Actieplan beter presteren* [Action Plan for Better Performance]. Retrieved from https://www.rijksoverheid.nl/documenten/kamerstukken/2011/05/23/actieplan-vo-beter-presteren.

Van der Kleij, F., Vermeulen, J., Eggen, T. & Veldkamp, B. (2013). *Leren van Toetsen: Een cyclisch proces [Learning from Tests: A cyclical process].* Arnhem: ToetsWijzer|Stichting Cito Instituut voor Toetsontwikkeling.

Voorstel Wijziging van de Wet op het voortgezet onderwijs, de Wet voortgezet onderwijs BES, de Wet college voor toetsen en examens alsmede de Wet op de expertisecentra in verband met onder meer de invoering in het voortgezet onderwijs van een leerlingvolgsysteem, een diagnostische tussentijdse toets en verplichte deelname aan internationaal vergelijkend onderzoek (leerlingvolgsysteem en diagnostische tussentijdse toets voortgezet onderwijs). [Amendment to the Secondary Education Act (BES), the Tests and Examinations Board Act, and the Expertise Centers Act in connection with such aspects as the implementation in the secondary education system of a student-tracking system, a Diagnostic Educational Test, and compulsory participation in an international comparative study (student-tracking system and diagnostic educational testing in secondary education)]. (2013, June 19). Retrieved from https://zoek.officielebekendmakingen.nl/kst-33661-4.html

2 Task construction

Wilma Vrijs and Sanneke Schouwstra

2.1 The task construction process

The construction of tasks for a formative, diagnostic assessment using the evidence-centered design (see Chapter 1) requires working methods that differ from those needed for creating summative tests (e.g., examinations). For the assessment experts who developed the DET, this difference was manifested primarily in the following components:

- **A student model**, which formed the foundation of the assessment: a model containing descriptions of all aspects and sub-aspects of the skills to be assessed. The student models for the three subjects of the DET are described in the Assessment Specifications (College voor Toetsen en Examens, 2014).
- **A task model** served as input for the context in tasks. More authentic interactions (task types) were applied, and innovative types of tasks were developed as needed, in order to allow more direct and authentic questioning. The Assessment Specification includes a description of the types of tasks appearing in the DET (College voor Toetsen en Examens, 2014).
- **The coding** of the response space (all possible responses), based on the student model. Coding is not about a correct/incorrect score for each task. The coding process requires identifying the skills and sub-skills to which a task refers and determining which student responses will have which diagnostic value.

Four stages can be distinguished in the task-construction process. This process began with the development of prototypical tasks during the preliminary study (Roelofs & Schouwstra, 2012). The assessment experts were supervised by educational researchers, particularly in the first stage. After the preliminary study, the second stage involved testing a selection of task types in small-scale try-outs (Cito, 2013a; Cito, 2014). This provided the assessment experts with a picture of which tasks functioned well in relation to the student models. This was followed by the large-scale task-construction process that is described in this section. After the large-scale task-construction process, the fourth stage involved additional construction (twice), for purposes of refreshing the item bank (the so-called seed items).

2.1.1 Construction assignments

The large-scale task-construction process was based on the construction assignments from the Board of Tests and Examinations (in Dutch, *College voor Toetsen en Examens*, or CvTE). The construction assignments of the Board of Tests and Examinations contain descriptions of *how many* tasks (responses) should be constructed for each subject, educational stream and aspect (or sub-aspect) in order to arrive at an accurate diagnosis concerning whether the student's command of the specific aspect is below, at or above the desired level. The *content* of the item bank is based on the student models for each subject, as recommended by the Assessment Specification committees. In this context, the term "task" refers to an assessment question on one screen. A task can consist of multiple responses.

A schematic overview of the number of responses to be constructed for each subject/skill is presented in Table 2-1 (including a margin for tasks eliminated after the pre-test and a limited number of tasks for an example assessment in each subject). The goal was to develop tasks for each sub-aspect within each educational stream that would yield 24 responses. To be able to give advice to students that a higher level educational stream or a lower level educational stream might be more suitable (so-called upstream or downstream indications) as well as for the standard setting, the construction assignment assumes an overlap of 50% between the item banks for adjacent educational streams. The overlap achieved during administration can be higher or lower, depending on the number of available tasks. For example, the vmbo-bb (the basic vocational program) tasks were administered to the vmbo-bb students, along with a portion of the vmbo-kb (the middle-management vocational program) tasks.

The numbers of responses in the construction assignments (from 2013, 2014, and 2015) have been calculated by multiplying the number of sub-aspects in the student model by the number of responses for the educational streams, including an overlap of 50%. For example, the calculation for English Writing Skills for vmbo (preparatory secondary vocational education) was as follows:

9 sub-aspects x 48 responses (12 suitable for vmbo-bb, 12 suitable for vmbo-bb and vmbo-kb, 12 suitable for vmbo-kb and vmbo-gt, 12 suitable for vmbo-gt) = 432 responses

Table 2-1. Overview of the numbers of tasks to be constructed according to the construction assignments

Subject	Educational stream	Responses according to the construction assignment for large-scale development	Responses to be generated each year for purposes of refreshing
Dutch Writing Skills	vmbo	576	115
	havo/vwo	468	94
Dutch Reading Comprehension	vmbo	528	106
	havo/vwo	396	79
English Writing Skills	vmbo	432	86
	havo/vwo	324	65
English Reading Comprehension	vmbo	672	134
	havo/vwo	504	101
Mathematics	vmbo	576	115
	havo/vwo	540	108

After completing the item bank, tasks were constructed each year for purposes of refreshing the item bank, in accordance with the 2015 construction assignment of the Board of Tests and Examinations. Item banks must always be refreshed because, in time, the tasks (and particularly the context of the tasks) become outdated, and because tasks are needed that correspond to the changing life experiences of the students. Replacing a number of tasks each year makes it possible to spread out the construction and analysis of tasks. Refreshing and expanding the item bank is also necessary for maintenance and in order to increase the quality, efficiency, and usability of the diagnoses and assessment (see the Quality Guidelines in the following section).

In the last year of the task-construction process, a series of example tasks was developed for each subject. The objective was to provide illustrative tasks to students and teachers for the various diagnoses offered by the DET. For example, a student performs above level on a sub-aspect. In that case, the task is illustrative of all tasks that make a good distinction between students performing at, and above level on the relevant sub-aspect. In this way, for each subject, educational stream, and sub-aspect, tasks were developed to be illustrative of a diagnosis of below/at level, and for a diagnosis of at/above level.

2.1.2 Phases in large-scale task construction

A subject team of about five assessment experts was available for the large-scale task construction. Most of these assessment experts were subject specialists with classroom experience, and with specializations in vmbo or havo/vwo. They provided internal supervision of the process: the realization of the numbers of tasks within the established timelines, in addition to collaborating in assembling the assessment, and the writing of scientific reports. The assessment experts also advised the confirmation committees of the Board of Tests and Examinations with regard to the tasks during the confirmation process. The subject teams were managed by a task-construction project leader, in addition to being supervised by a R&D project leader, and an ICT project leader.

Under the supervision of the subject teams, construction groups (consisting of teachers) started with the creation of basic tasks. These tasks were then screened and improved multiple times, based on the findings of the screening. After the tasks were approved by the subject teams, they were submitted to the confirmation committee of the Board of Tests and Examinations. The confirmed tasks were then pre-tested,

in order to determine the quality of the tasks using empirical data, after which the tasks could be calibrated. The pre-test was followed by the final confirmation procedure for the adaptive administration of the DET.

Basic tasks created by construction groups

Construction groups were used for the development of the basic tasks. These groups consisted of classroom teachers from the various educational streams and years for which the DET is intended. This ensured that the tasks would correspond to the level and life experiences of the students. All of the subject teams with assessment experts opted to have two separate construction groups: one for vmbo and one for havo/vwo. The teachers from the construction groups were supervised by the assessment experts. The assessment experts and teachers met regularly in order to discuss specific tasks and the working methods in general.

The teachers in the construction group were trained through a one-day general course in assessment techniques, and a customized half-day program focusing specifically on the DET and the operationalization of the student model for the relevant subject. The teachers were subsequently given a development assignment based on the student model. For example: Create five tasks on *Tuning to audience and objective*. The teacher submitted a first draft of a task to the assessment expert. The assessment expert provided feedback, and the teacher adjusted the task. A second draft was submitted, and the assessment expert reviewed the task again. On average, a teacher created approximately five tasks per week, for which purpose the teacher had about half a day per week available.

To create the basic tasks, the teachers used a special Word template designed especially for this purpose. This template helped them in the selection of a type of task (see Figure 2-1) and in the entry of all of the necessary metadata for the task. The templates also made it possible to create new forms of tasks, long before it was technically possible in the authoring environment and test administration player. The Word templates also simplified the work of the assessment-support staff, who entered the tasks into the item bank, as all of the tasks were delivered in a standardized form. The teachers did not enter the tasks themselves, as this task required technical training and experience with the authoring environment.

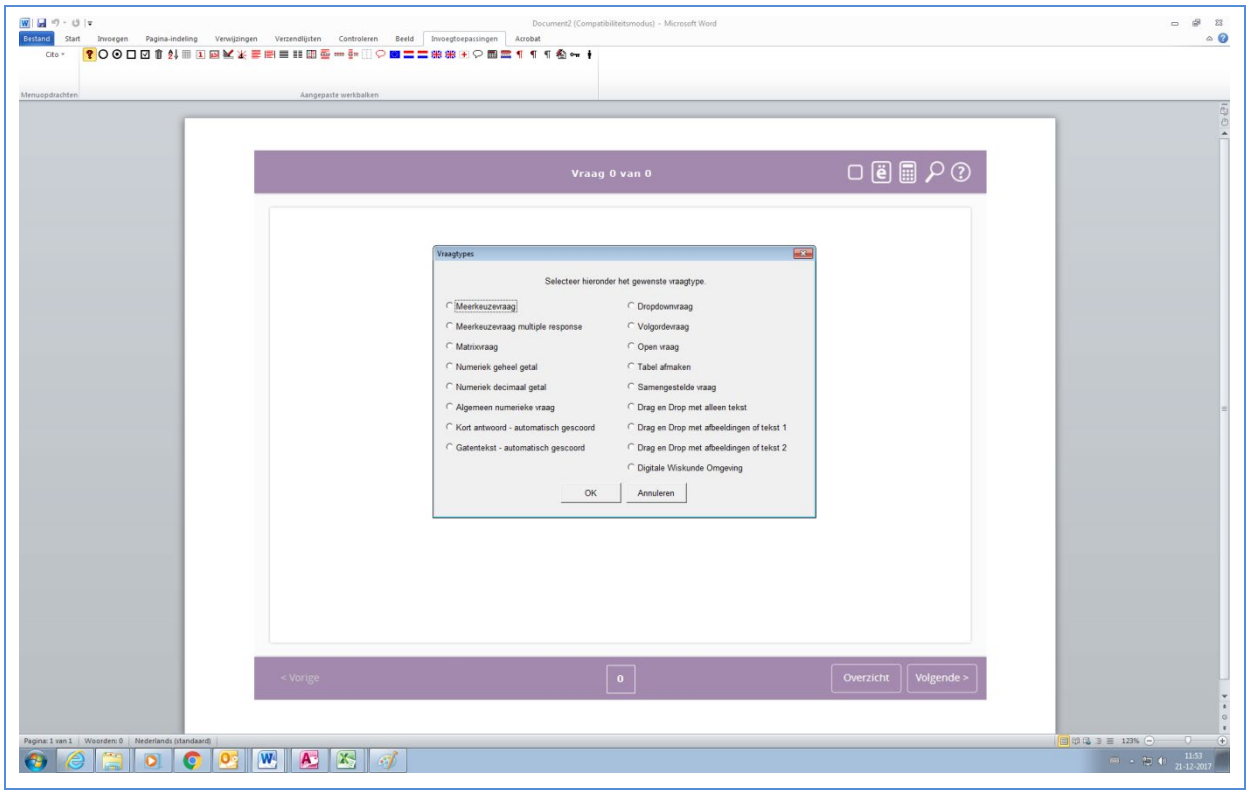


Figure 2-1. The choice of a task type in the Word template

Screening

Once the assessment expert was satisfied with the tasks, other assessment experts screened them. The process of screening consisted at least of the following:

- A subject-based content screening by the assessment experts from the relevant subject team.
- A technical assessment screening, also by assessment experts from the subject team.
  A “cross-wise” screening was used for the subject-based content screening and the technical assessment screening. In this option, the assessment experts specialized in havo/vwo screened the vmbo tasks, and vice versa. This ensured a connection between the educational streams.
- A screening for language usage/Dutch by a department responsible for the logistical production of tests within the framework of quality management.
- A screening (by the same department) for compliance with agreed-upon guidelines concerning layout, and language usage. E.g., the source must always appear on the left, and the question should always appear on the right. The question is always printed in boldface.
- Technical screenings by the assessment experts and staff of the Operations Office, for the purpose of assessing whether the separate assessment packages function completely as they should on such aspects as operation (e.g., does the dragging task work?), scoring/coding (e.g., is the correct answer calculated correctly?), etc.

During the process, the task was entered into the software application for creating tests and managing item banks (the authoring environment Questify Builder). One of the screenings was conducted in the authoring environment, and a final technical screening was conducted in the digital administration system (Facet). The goal was to review the task one last time, exactly as the student will see it.

Quality Guidelines

During the development of tasks and during the screening, attention was paid to the following quality aspects, which were formulated in the construction assignment (College voor Toetsen en Examens, 2013).

Box 2-1. Aspects of quality as formulated in the construction assignment (College voor Toetsen en Examens, 2013).

Validity and authenticity

To which extend does the instrument capture the intended aspect of the student model? This includes evaluating whether other sub-skills or variables unintentionally play an important role in solving a task correctly.

Reliability, difficulty and discriminating power

Isn't the task too difficult or too easy for the intended program/track? Does the task discriminate well between students who are below level, students who are at level and students who are above level? As a rule, tasks that nearly all students answer correctly or, in contrast, that nearly all students answer incorrectly, are hardly informative and discriminative.

Diagnostic information value

To which degree are the tasks and response categories informative about the mastery of one or more sub-aspects of the student model? Can all possible answers and response categories be linked adequately to the level of mastery of the intended sub-aspects?

Efficiency

Isn't too much response time needed for the task, given the time available for taking the entire assessment?

Functionality of the source material and the task situations

- Is the source material functional and necessary, given the purpose of the measurement?
- Is the diversity of the source material and task situations sufficient? A task situation is the context of the task. For example, for a writing task one can think of filling out a form or writing a letter or writing report.
- Is the source material recognizable for the student and not too complex or too extensive?
- Is the source material admissible?

Confirmation

Tasks that had been entered into the authoring environment and that had gone through all of the screenings were submitted to a confirmation committee. The Board of Tests and Examinations established such a committee for each subject: Dutch, English, and mathematics. Each confirmation committee consisted of four classroom teachers from the educational streams and years for which the DET is intended, along with a chair. The members of the committee were tasked with determining whether tasks or complete assessments were approved for administration.

Several confirmation days were planned throughout the year. For each confirmation day, the subject teams provided a set of tasks to the confirmation committee for approval. The tasks were presented electronically through the administration environment used by the schools (Facet), such that the members of the confirmation committee were able to view all of the tasks digitally in the same way that they would be presented to the student.

The confirmation committee assessed each task and provided suggestions for improvement as needed. In the meeting, the tasks were discussed and either approved or rejected. In some cases, adjustments were suggested for tasks that had been rejected, so these tasks could be approved as well later on. In its assessment, the confirmation committee considered the following and other aspects:

- The task's level of difficulty: Is the level of difficulty appropriate to the intended educational stream?
- The operationalization: Is this a good way to measure the relevant sub-aspect of the student model?

- The alternation within the set of tasks and within aspects and sub-aspects: Is there sufficient diversity in the type and content of tasks?

In the authoring environment, the assessment experts kept a record of the tasks that had been approved. Regular attention was paid to whether sufficient material had been approved in order to achieve the numbers of tasks necessary for accurate diagnoses in the construction assignment. If needed, more tasks could be constructed.

Pre-testing, analysis and adaptive administration within the test-construction process

After the confirmation, a pre-test design for all of the tasks was created (Cito, 2015, Chapter 3; Cito, 2016,Chapter 3), and it was reviewed at a substantive level by the assessment experts. The tasks were then pre-tested in two successive years.

After the pre-test, test and item analyses were conducted in order to evaluate the quality of the tasks (Cito, 2015; Cito, 2016). The assessment experts evaluated the results according to several criteria (see Chapter 10 on Data processing and item analyses). All tasks for which more than half of the responses were appropriate with regard to the level of difficulty, discriminating power, and relation with the other responses for the same sub-aspect were automatically approved. The content of all tasks that were potentially problematic were inspected by the assessment experts and discussed with the confirmation committee, which determined whether a task would be approved for inclusion in the adaptive test.

2.2 Task types in the DET

Efforts were made to develop more authentic tasks for the DET, in order to allow more direct assessment. This is important for diagnostic tasks, in order to allow for the accumulation of evidence about the student's level of skill using the aspects of the student model (Evidence Centered Design). Considerable diversity in types of tasks was created, and this made it possible to appeal to the skills of students in a direct manner and to offer students a varied assessment in which to demonstrate their skills (see Figure 2-2). In authentic assessments, an attempt is made to have the form and content of the assessment correspond as closely as possible to the manner in which students would apply the skill that is being measured, in real-life situations. According to Frey, Petersen, Edwards, Pedrotti and Peyton researchers are increasingly emphasizing the importance of authentic assessments (2005). In addition, digital assessments make it possible to offer more complex, dynamic and construct-relevant tasks (Jodoin, 2003). Moreover, as Cheng and Basu argue in their article of 2006 on the use of multimedia in tasks, innovative task forms can increase the motivation and involvement of students (quoted in Goosen & Vernooij, 2017).

Most task types can be used for all subjects, and a few are specific for languages or developed specifically for mathematics. In most types of tasks, task constructors can add text, images, audio, video, and mathematical expressions. For a few types, images are used as a base for the interaction to be constructed. The various interactions can be divided roughly into clicking, dragging, and filling in (typing).

Based on the student model and the task model, the assessment expert determines which type of task is most suitable for measuring the relevant aspect. In some cases, multiple options are possible. Several task types were developed especially for the DET, in order to allow a more direct assessment. For languages, this includes a paragraph task (see Figure 2-3) and the task type in which the student can click on text elements (e.g., find the word with the same meaning, find the spelling error, and click on the signal word that provides evidence of this). A type was also developed in which the student is able to both mark text and make corrections or other changes: the correction task (see Figure 2-4).

For mathematics, various types of tasks were developed in which the student is able to perform various mathematical interactions. Examples include drawing a chart or drawing or adjusting a geometric figure (see Chapter 8 on Automated assessment of mathematics). These special types of tasks were developed with input from the assessment experts.

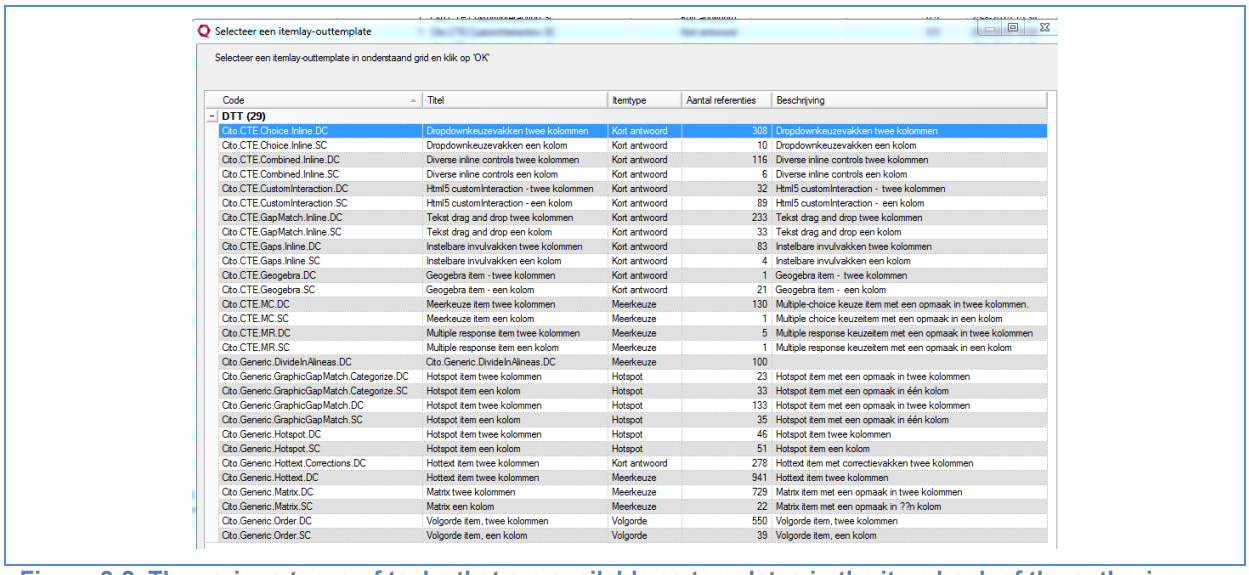


Figure 2-2. The various types of tasks that are available as templates in the item bank of the authoring environment (described in section 2.2.1)

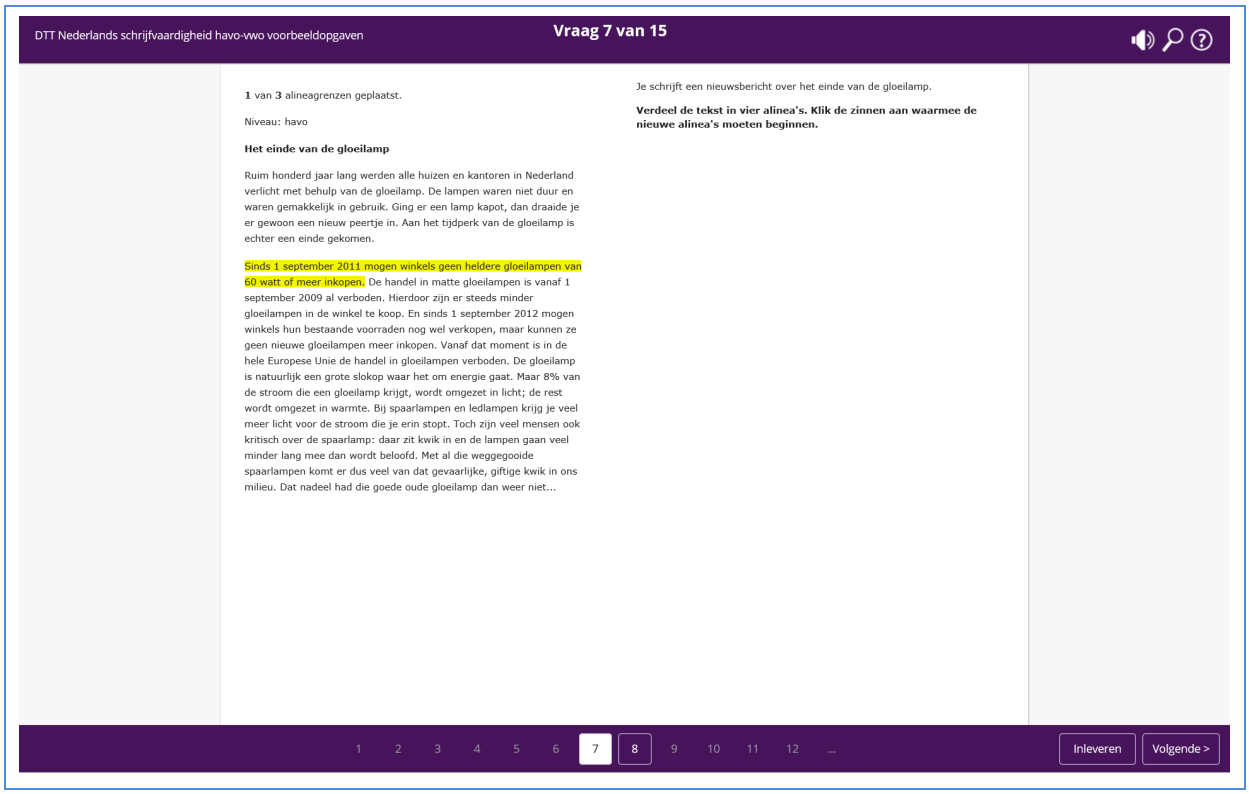


Figure 2-3. Example of a paragraph task in which the student has to divide the text into four paragraphs. The student should click the first sentence of each new paragraph.

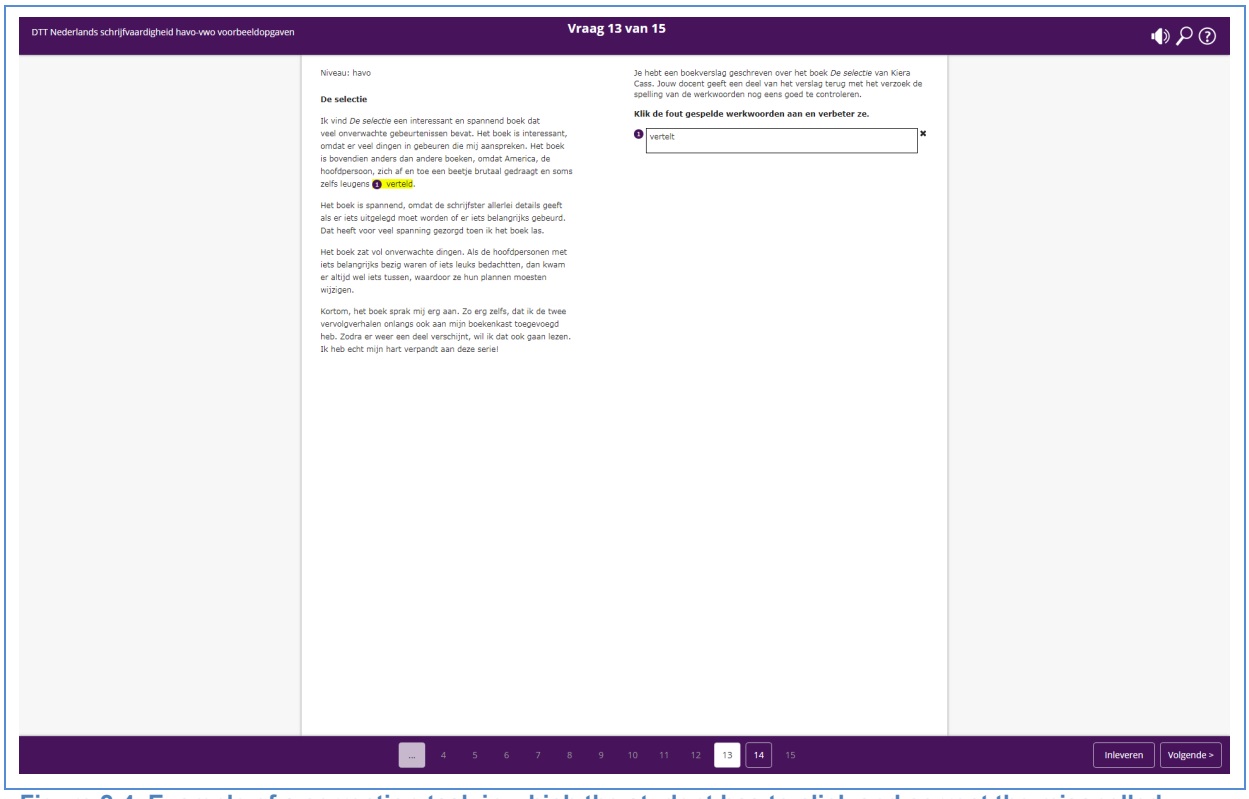


Figure 2-4. Example of a correction task in which the student has to click and correct the misspelled words

2.2.1 Overview of digital diagnostic task types

All of the types of tasks used within the DET are included in the following table, along with concise descriptions. Examples (in Dutch) of each type of task are presented in Appendix 17.1.

Table 2-2. Task types used in the DET

Type of task	Description	Interaction
Paragraph task	A type of task in which the student divides a text into paragraphs. The paragraph divisions are visible in the text. Suitable for allowing the student to apply text structure.	Clicking, multiple choice
Categorization task	The student drags particular elements (text, images) in fields by category. Suitable for sorting elements (e.g., formal/informal for tasks concerning types of text).	Dragging
Combination task	A type of task in which multiple types of interactions (e.g., multiple choice and open-ended task) are combined on a single screen.	Miscellaneous
Correction task	The student selects and improves components of a text. Suitable for assessing grammar and spelling.	Clicking and entering
Drop-down task	A type of task in which the student selects an answer from a list of options. The student immediately sees the effect of the choice in the text. This type of task is often used for vocabulary questions in which the student inserts the most suitable word into the text.	Clicking, multiple choice
DME task (task using the Digital Mathematics Environment)	The student can draw chart points or charts as an answer.	Miscellaneous
GeoGebra task	The student can perform various mathematical interactions as an answer. Example: drawing or adjusting the correct geometric figure.	Miscellaneous
Hotspot task	The student clicks on the correct answer in an image.	Clicking
Short open-ended task	The student types in an answer. This type of task is used for languages, in order to test spelling. For mathematics tasks, students can use a special formula editor to enter input in the form of numbers and formulas.	Entering
Marking task	The student selects active parts of a text. Suitable for finding answers in a text (e.g., a quotation or misspelled word).	Clicking
Matrix task	The student selects with several questions in a row one of (usually) two options (e.g., true or false). Suitable for asking about several details.	Clicking, multiple choice
Multiple-choice task	The student selects the correct answer from a list of alternatives (texts, images, videos, audio fragments) or from active areas in an image.	Clicking, multiple choice
Multiple-response task	The student selects one or more answers from a list of alternatives (texts, images, videos, audio fragments). This form is often used for tasks involving internet forms (then it is actually a kind of sorting question).	Clicking, multiple choice
Dragging task	The student drags an answer in the form of text. Suitable for matching or sorting/categorizing.	Dragging
Drag task with image	The student drags an answer. Suitable for linking images to words.	Dragging
Ordering task	The student drags text fragments or numbers in a new order.	Dragging

2.2.2 The choice of a task type

As previously stated, the student model and the task model constituted the foundation for a task. This was followed by considering which type of task would be most suitable. The following are several considerations when selecting the most suitable type of task:

- Select the form of interaction that will allow the student to demonstrate the skill or sub-skill to be assessed as authentically, directly, and efficiently as possible.
- Consider efficiency for more complex interactions (e.g., ordering tasks): Would this task take too much time relative to the information that it yields?
- Various types of tasks are possible in which students immediately are able to see the effect of the interaction in the text (e.g., by entering an answer from a list of options, dragging an answer into a field, or typing an answer). It is more authentic when the effect can be seen, but it is often easier as well.
- A specific characteristic of short open-ended tasks is that students enter the answers themselves. In general, this is more difficult than selecting or dragging a word from a list.
- Ensure a good balance between alternation for the student and uniformity of style and interaction.
- The DET is unique in the manner in which the tasks are automatically "scored." Simultaneous consideration of the task and its "scoring" promotes the consistency of the tasks.

As much as possible, the assessment experts examined which type would be the most authentic and suitable. The first try-out and pre-test generated some experiential information (Cito, 2013a; Cito, 2014; Cito, 2015; Cito, 2016):

- If the student must arrange text (or other) elements in an order – as with chronological events (text comprehension) or numbers from low to high, an ordering task would be a good choice. Although such tasks do require more time, they are a direct manner of questioning.
- If the student must categorize text (or other) elements (e.g., formal/informal language usage or prime number/not a prime number), a categorization task would be a good choice.
- Aspects like spelling can be assessed through free text fields or through types of tasks in which the student can correct spelling errors in the actual text (correction tasks).
- Mathematics tasks in which students must type in their own answers are more difficult than multiple-choice questions.
- If the student must demonstrate understanding of text structure, a good choice would be a paragraph task, in which students divide texts into paragraphs.
- The design of a task can also contribute to authenticity: if the student must complete a form, be sure that the task looks like a form.

Task characteristics and coding

When constructing a task, the constructors had to consider the concept to be assessed and the evaluation, as well as the coding of the link between the (almost) correct answer and the related aspect of the student model. For example, in a multiple-choice task the constructor should be able to indicate that the correct answer D provides information about the sub-skill 2.4 from the student model for Dutch writing skills. The possibility for coding the tasks like this has been developed especially for the DET. In the next chapter (3), the coding is discussed extensively.

The following should also be stated for each type of task:

- The aspect of the student model that is being surveyed
- The educational stream for which it is intended
- Which interim objectives are covered by the task
- For mathematics, the type of scoring is indicated

In addition, several characteristics were important from the perspective of item bank management, including the pre-test in which the task was included and when it first appeared in the adaptive test.

2.3 In conclusion

The development of large quantities of diagnostic innovative task types required a lot of creativity and flexibility from the assessment experts, especially given the tight time frame for development. Nonetheless, they have succeeded in developing a diverse assessment instrument that resulted in authentic questioning

and proved to yield a reliable result. In the following chapters, the development is discussed for each subject separately.

2.4 References

Cheng, I., & Basu, A. (2006). Improving multimedia innovative item types for computer based testing. In *Eighth IEEE International Symposium on Multimedia*, 557-566. doi:10.1109/ism.2006.92

Cito (2013a). *Diagnostische tussentijdse toets: Verslag try-out 2013 [Diagnostic Educational Test: Try-out report 2013].* Arnhem: Cito.

Cito (2014). *Diagnostische tussentijdse toets: Verslag try-out 2014 [Diagnostic Educational Test: Try-out report 2014].* Arnhem: Cito.

Cito (2015). *Diagnostische tussentijdse toets: Verslag pretest 2015 [Diagnostic Educational Test: Pre-test report 2015].* Arnhem: Cito.

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016 [Diagnostic Educational Test: Pre-test report 2016].* Arnhem: Cito.

College voor Toetsen en Examens (2014), *Publieksversie Toetswijzer Diagnostische Tussentijdse Toets voor Nederlands, Engels en Wiskunde [Public Version of the Assessment Specification for the Diagnostic Educational Test for Dutch, English and Mathematics].* Utrecht: College voor Toetsen en Examens. Retrieved from https://www.pilotdtt.nl/documenten/publicaties/2014/12/15/toetswijzer-dtt

Frey, B. B., Petersen, S., Edwards, L., Pedrotti, J. T., Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357-364. doi:10.1016/j.tate.2005.01.008

Goosen, R. & Vernooij, S. (2017). *Innovatieve vraagvormen in de DTT schrijfvaardigheid Engels: Een onderzoek naar de geschiktheid van meer authentieke vraagvormen. [Innovative question forms in the DET for English writing skills: A study of the suitability of more authentic types of questions].* Unpublished internship report. Arnhem: Cito.

Jodoin, M. G. (2003). Measurement efficiency of innovative item formats in computer-based testing. *Journal of Educational Measurement*, *40*(1), 1-15. doi:10.1111/j.1745-3984.2003.tb01093.x

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

3 Response processing: Coding instead of scoring

Joke Hofstee

The DET is adaptive, thus allowing making many diagnoses within the testing time allowed (see Chapter 1). In order to arrive at reliable diagnoses within the available testing time, it is important for the tasks in an assessment to be efficient. In other words, a task should provide as much information as possible while requiring as little time as possible for answering. In addition, for purposes of validity, it is important for tasks to be authentic. In other words, the task that the student must perform in order to answer the task should provide the closest possible reflection of the skill that the assessment is intended to measure, while measuring as few non-relevant skills as possible. Consider a small example: inserting blank lines in a text to separate paragraphs is substantially more authentic than typing in the numbers referring to the lines between which a paragraph separator should be inserted. In order to create an item bank containing tasks that are as efficient and authentic as possible, it was necessary to develop new types of tasks, as described in the previous chapter (see Chapter 2).

In addition to the content of the tasks, there was a need to implement a new manner of response processing. A task can consist of multiple parts (interactions), each of which can have its own links to aspects of the student model. For each part of a task, it is important to indicate which answers are correct and how these answers are linked to the student model. In the DET, this linking of answers to the various aspects of the student model is known as coding. One difference between coding and scoring is that coding is not simply summed in order to produce a final score or a sum score for each sub-aspect or main aspect. Instead, it is used to calculate the probability that a student is performing below, at, or above level on each aspect (discussed in Chapter 12). In this calculation, the answers to the task parts are not necessarily taken together. They can also be used as separate observations. This enhanced the efficiency of the DET.

The following sections discuss how this manner of coding is constructed and designed in the authoring environment, as well as their influence on the working methods used in the construction of new tasks.

3.1 Concept definition: Items and responses

Although the concept of “items” is used in both the literature on testing and in the psychometric literature, it is not always understood in the same way. The concept of “items” is also used in the QTI (“Question & Test Interoperability”; see Chapter 9) standard, which is adopted in the authoring environment and the test administration player, Facet.

A simple multiple-choice task is understood as being an “item” for assessment experts, psychometricians, and in the authoring environment. For example, for the item in Figure 3-1, only one interaction is possible: the choice of an alternative. After the task has been answered, only one response will be assessed as correct (1) or incorrect (0): there is a single outcome. The concept of “item” is less clear for more complex tasks with multiple components that could lead to multiple responses and to one or more outcomes.

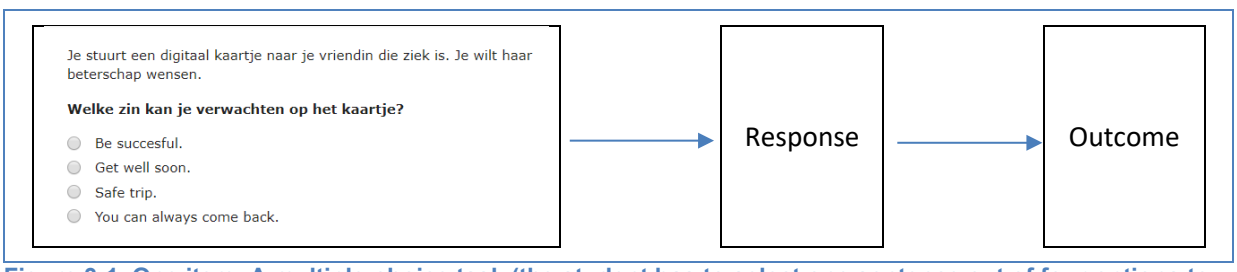


Figure 3-1. One item: A multiple-choice task (the student has to select one sentence out of four options to put on a postcard)

For example, in Figure 3-2 you can see a task consisting of three components. In the QTI standard, this is an item with three interactions[4]. In this example, each response can be assessed as correct or incorrect, independently of the other responses and regardless of whether the other responses are correct or incorrect. For psychometricians there are thus three item responses.

In the remainder of this chapter, the term “task” is used for a question that is presented on one screen and is a substantive unity.

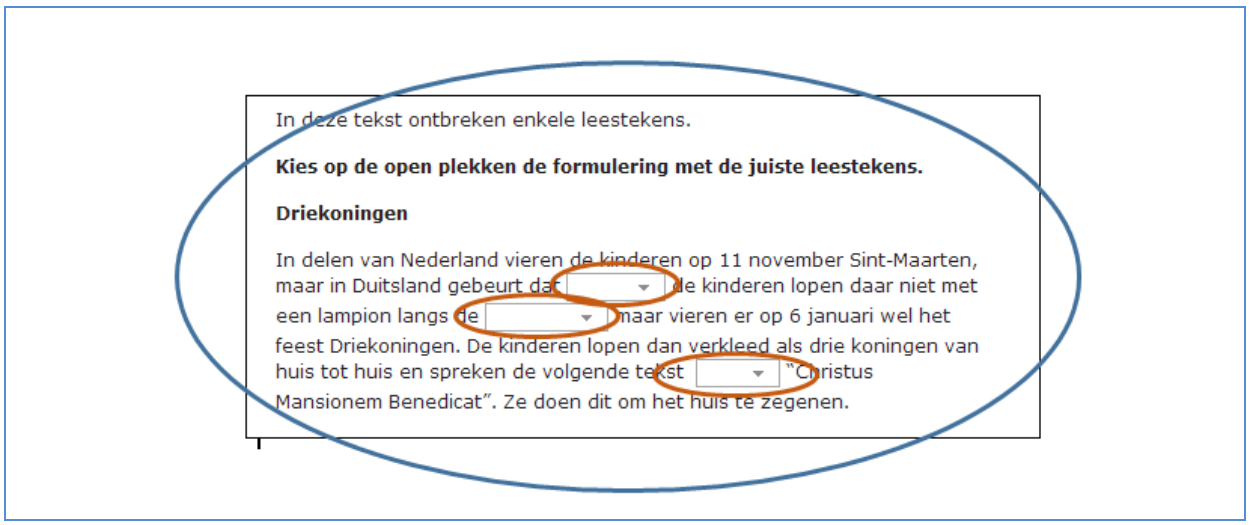


Figure 3-2. Illustration of item definitions in QTI: one item (blue oval) with three interactions (red ovals). In this task, the student has to insert a punctuation mark using a drop-down menu at three places in the text.

3.2 Grouping of interactions in the authoring environment

The authoring environment in Questify was modified for the DET, such that responses could be evaluated and linked to one or more main aspects and sub-aspects of the student model both separately and in groups. For each task, the constructor had to determine whether responses should be evaluated separately or whether they should be evaluated as a whole.

In general, responses with internal dependencies should be evaluated as a whole. For example, in a task in which the student must select both a number and a scale, the number (e.g., 100) cannot be evaluated independently from the selected scale (e.g., “cm”). For this reason, the score editor of the authoring environment (see Figure 3-3) offers the option of grouping responses.

[4] In QTI the names in full are ‘assessmentItem’ and ‘interaction’

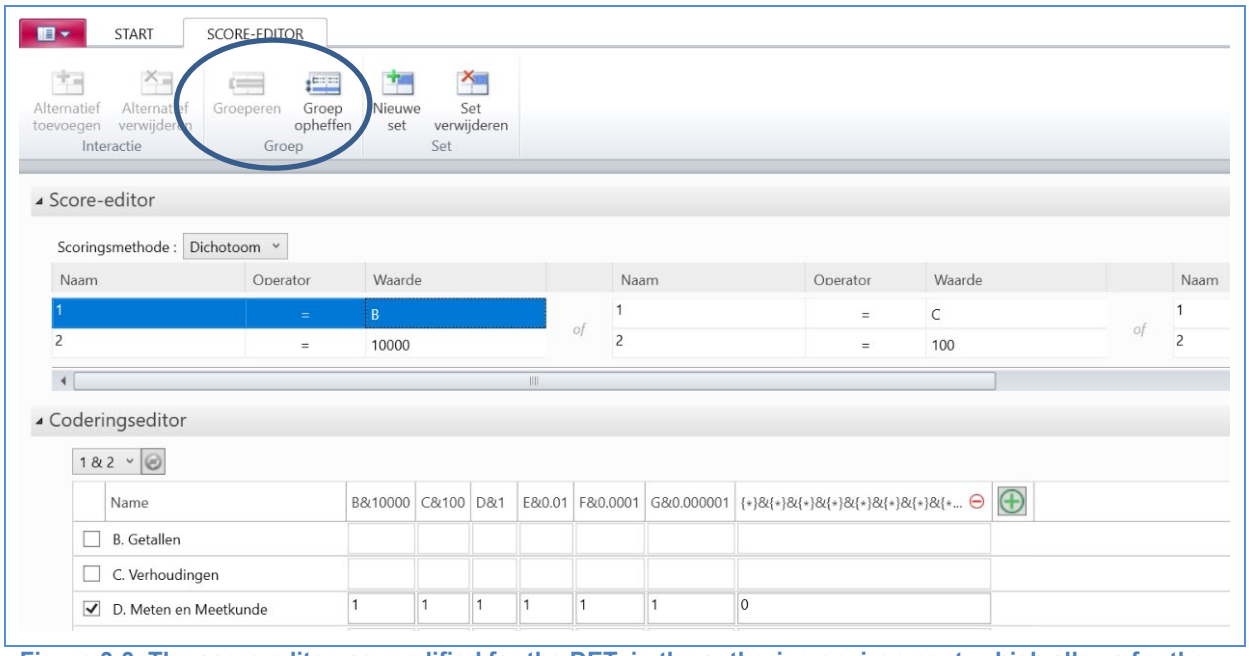


Figure 3-3. The score editor, as modified for the DET, in the authoring environment, which allows for the grouping of responses

An example in which the consideration of whether responses should be evaluated separately or together is more difficult is presented in Figure 3-4. In this case, the considerations include the following:

- Is the substantive coherence between the interactions very strong? (No, the task is actually three separate questions.)
- How often can the response alternatives be selected: once, or more often? (If an alternative can be selected only once, there is always one less alternative after each response.)
- How many response alternatives are offered: more than the selections to be made? (In this case, more; each interaction remains a multiple-choice task)

In this example (see Figure 3-4), there are three independent responses, each of which can make an independent contribution to the diagnosis. For this reason, the responses have to be evaluated separately. Each of the three responses must subsequently be linked to the student model in Questify.

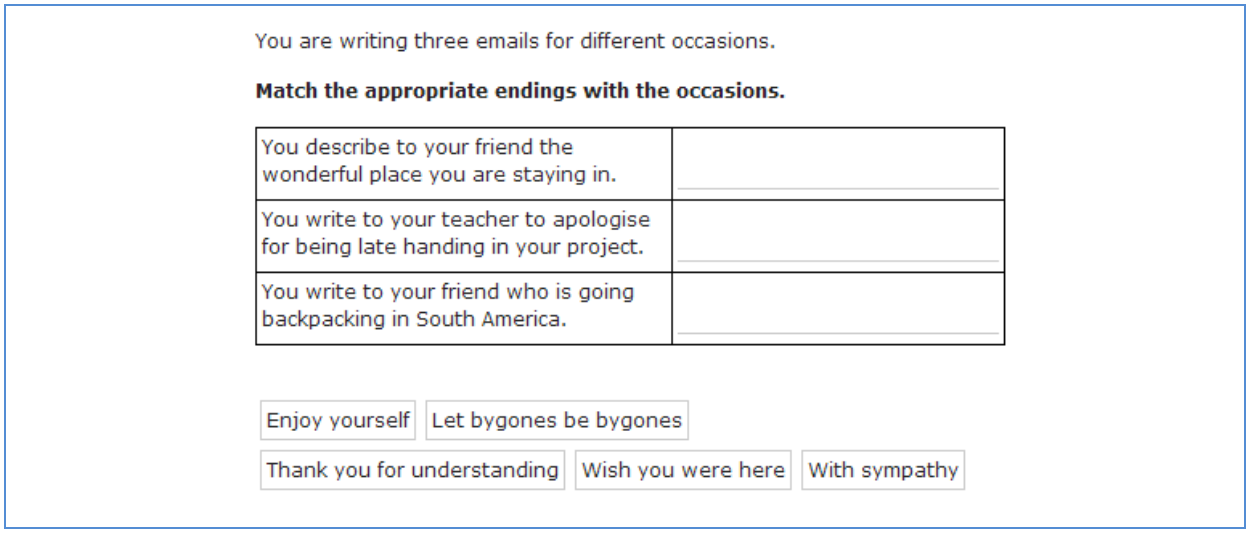


Figure 3-4. Example of a task with multiple interactions

3.3 The student model in the authoring environment

The student models in the DET describe multiple aspects for each skill. These aspects can be subdivided into main aspects and sub-aspects with a tree structure, as depicted in the following hypothetical example for English writing.

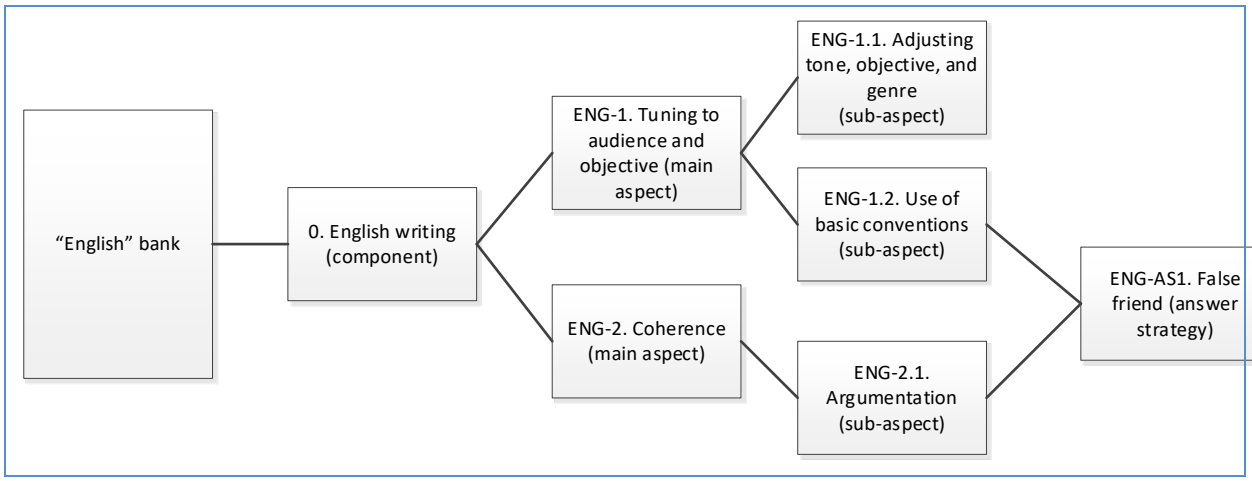


Figure 3-5. Example of a hypothetical tree structure for English writing

In addition to the main aspects and sub-aspects, it is possible to define answer strategies. For example, "False friend" is indicated in Figure 3-5. "False friend" is an incorrect answer strategy in which the meaning of the English word is confused with a word from another language (e.g., Dutch) that closely resembles it in terms of spelling or form. For example, consider the word "actual," which can be confused with the Dutch word "*actueel*," which means "current." For the DET, although answer strategies have yet to be described in the student models, they have been enabled in Questify. Defining of a complete student model[5] (as above), the connection with tasks and interactions, and the processing in an adaptive module have been enabled in the authoring and administration environments (see Chapter 9).

3.4 Linking of interactions to the student model in the authoring environment

In the authoring environment, the constructor must draw a substantive link between the student model and the various responses. In the coding editor that was developed, it is possible to indicate the aspect and sub-aspect on which the answer provides information. For example, in a multiple-choice task, it can be indicated that the correct answer is B. In the coding editor, it can then be indicated that the correct answer provides information about Main Aspect 1 ("Coordinating to audience and objective") and Sub-aspect 1.1. ("Coordinating tone and register"); see Figure 3-6.

[5] In the authoring environment, a student model is referred to as a concept structure, and the coding is referred to as concept coding.

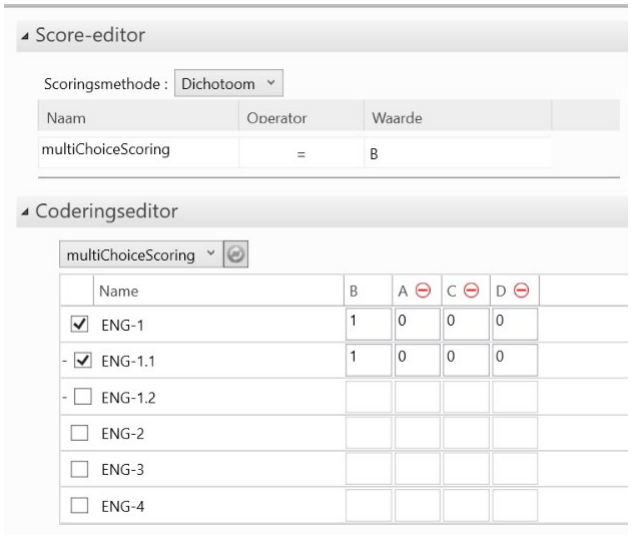


Figure 3-6. Coding of a multiple-choice task

The coding is more complicated for tasks with multiple interactions. The constructor must determine which answers are possible and what they indicate. Each interaction can lead to a separate response that is either correct or incorrect. Alternatively, the responses can be grouped and assessed together as either correct or incorrect. For each outcome, the constructor must determine the aspects and sub-aspects for which the student's answer contributes to the diagnosis. For example, in the ordering task in Figure 3-7, all of the responses have been grouped, together forming one outcome. Only one order can be considered correct (B, E, D, A, C), and all other orders must be evaluated as incorrect. In the coding editor we can see, that the correct order provides information on Main aspect 2 and Sub-aspect 2.1.

For interactions involving a large number of possible answers (e.g., in correction tasks the possible answers can be over the 100) applies that each of these possible answers must be categorized according to some sort of rule before they can be processed. This might mean that a number of the responses are grouped together, with only one outcome coded as correct and all other outcomes coded as incorrect. It might also mean that certain rules or pre-processing of the answer are used, e.g., deleting blank spaces, converting to lower case, or using a computer algebra system to determine mathematical equivalence. The coding editor was developed specifically to do this for the DET.

It also allows for designating multiple outcomes (orders) as correct. For example, in Figure 3-8, two different orders have been designated as correct, and all other orders are incorrect. It is also possible to divide the outcome space in more than two categories. For example, instead of "correct" and "incorrect," it could be divided into "completely correct," "somewhat correct," and "incorrect." In the example in Figure 3-9, the first order is completely correct (Category 2), the second order is somewhat correct (Category 1), and all other orders are incorrect.

The task of the constructor is thus expanded considerably. In addition to developing the tasks, the constructor must consider which answers are possible, whether answers must be processed before assessment, which answers should be assessed together, which answers are correct (or somewhat correct), and about which aspects of the student model the correct answers provide information. To enable all those tasks the standard score editor in the authoring environment was modified.

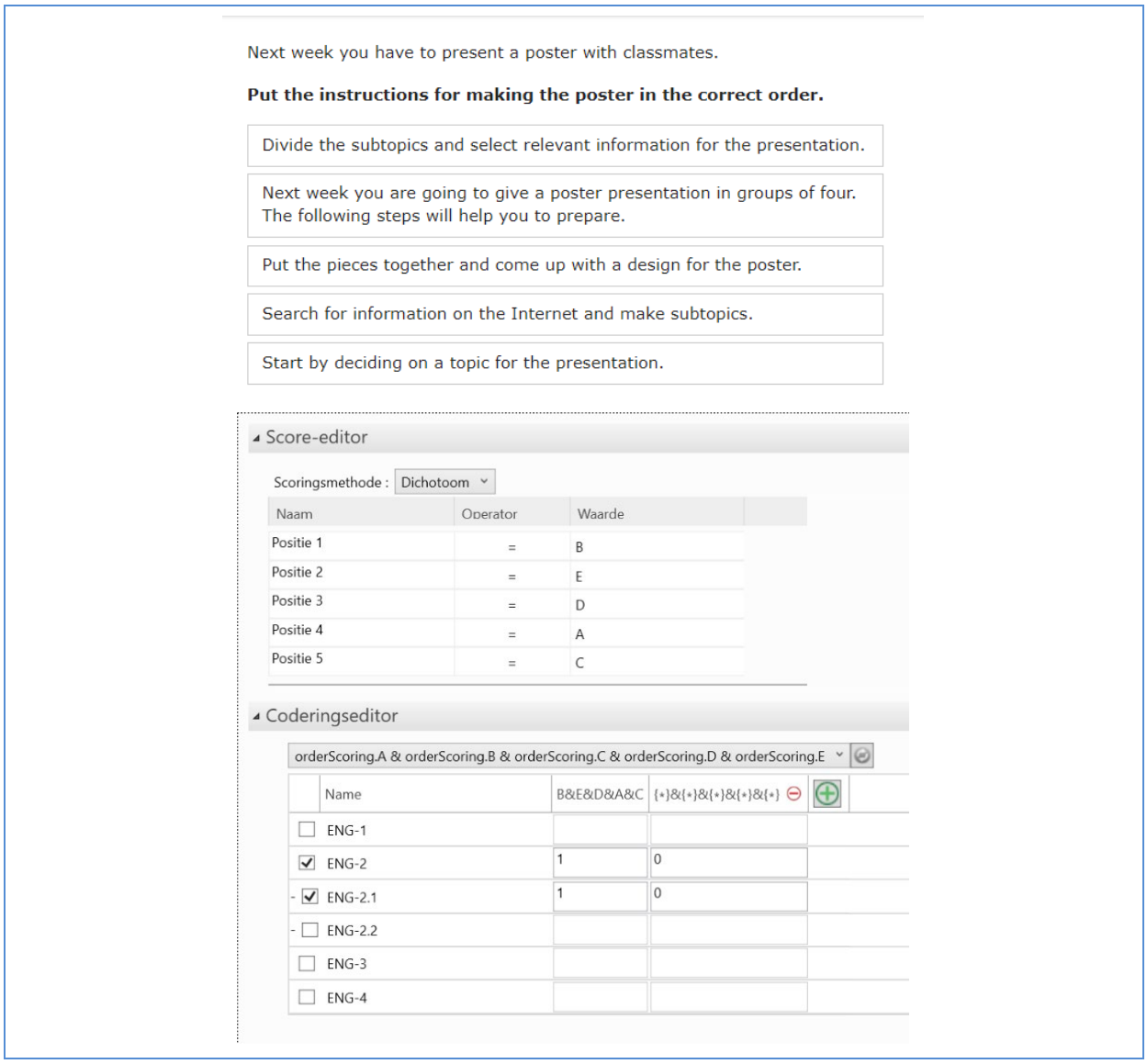


Figure 3-7. An ordering task in which only one order is correct

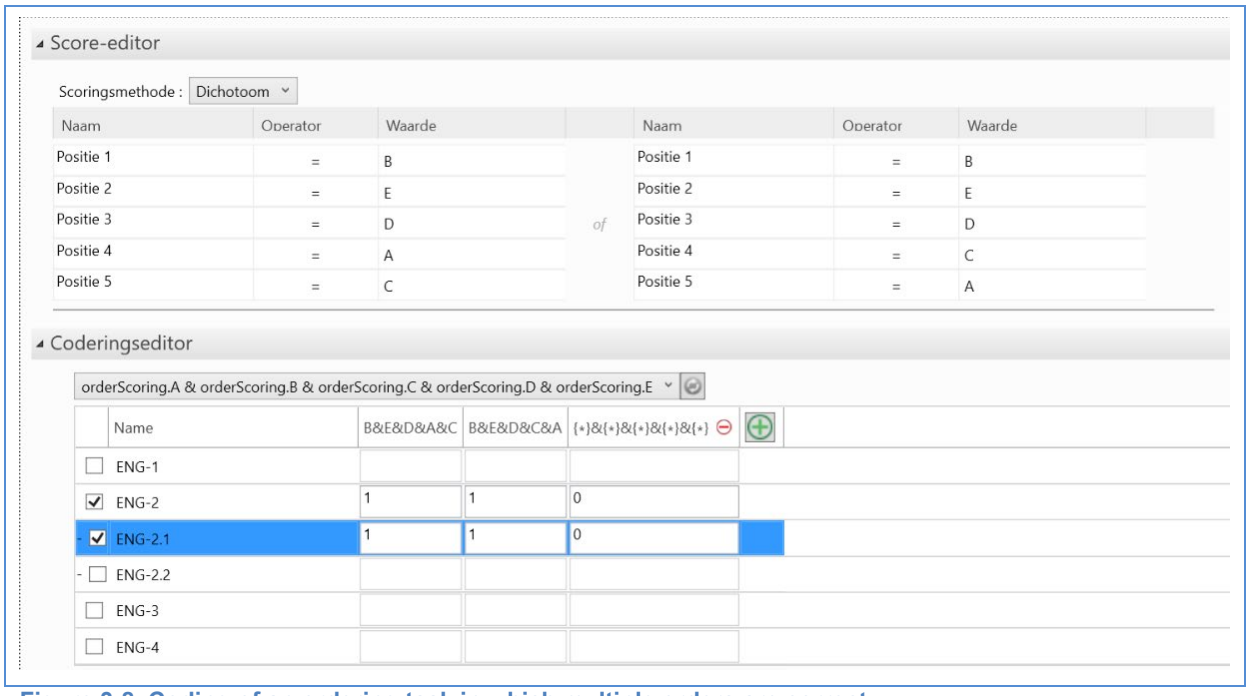


Figure 3-8. Coding of an ordering task in which multiple orders are correct

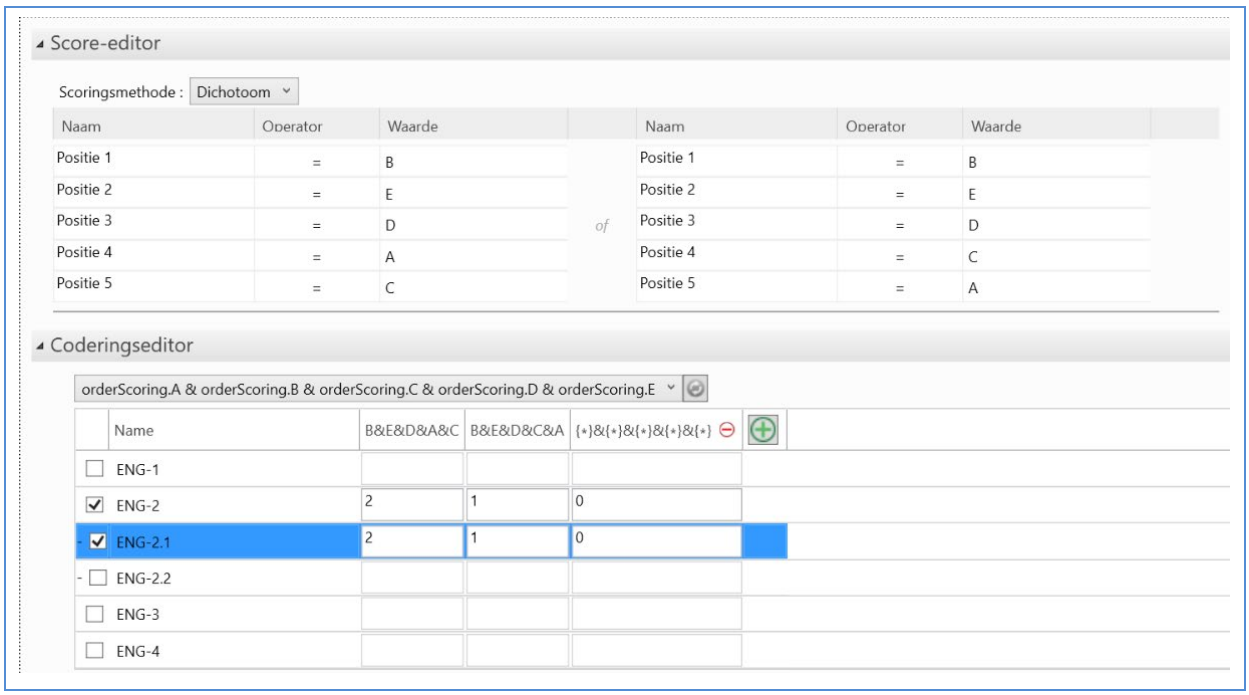


Figure 3-9. The coding of an ordering task in which one order is correct, another order is somewhat correct, and the rest are incorrect

4 DET Dutch

Uriël Schuurs, Kirsten van Ingen, Roelien Linthorst and Laura van Hofwegen

At the start of the project, an exploration was conducted to identify the most important skills and sub-aspects that should be included in a diagnostic test for Dutch in the lower years of preparatory secondary vocational education (vmbo), senior general secondary education (havo), and pre-university education (vwo). It was decided to start by exploring the possibilities of a diagnostic writing-skills test. The most important reason for not choosing reading comprehension was that, in practice, reading comprehension is the most tested skill and that the elaboration of this skill within the framework of the preliminary study was not expected to generate many innovative insights. In contrast, it is known that writing skills have not received sufficient attention in secondary education, and there are options for automated administration and assessment, although these options are not necessarily obvious. Digital assessment does not seem to be much of a problem for reading and listening. There was thus little reason to assume that it would be necessary to overcome problems relating to automation for these skills in the further course of this project. In contrast, speaking and conversational skills would be extremely difficult to test automatically within the project framework. The further course of the project did involve exploring the possibilities for the diagnostic assessment of Reading Comprehension, Listening, and Speaking. Tasks were developed and pre-tested only for Writing and for Reading. This chapter is therefore limited to Writing Skills and Reading Comprehension.

4.1 Dutch Writing skills

4.1.1 Student model for Dutch Writing skills

One important property of the student model that was selected for the DET is that it addresses characteristics of the writing product, while emphasizing aspects of the writing process. In the past two decades, studies have highlighted the complexity of writing skills, which – like language acquisition in general – does not display a linear build-up. Moreover, throughout the writing process, a writer is constantly working with a variety of interacting sub-processes, including planning the rest of the text, putting thoughts into words and revising previously written text fragments (Rijlaarsdam, et al., 2011). Skillful writers do this consciously as a part of their writing strategies. The student model for writing is linked to the model developed by a research group of the Educational Testing Service (Princeton), which provides a good summary of scientific insights into processes of reading and writing (Deane et al., 2008; Deane, 2011). Since 2007, this group has been developing formative and summative tests that generate information on the quality of both the writing products and the writing processes of students. For a detailed description of the model, see the report of the DET preliminary study (Roelofs & Schouwstra [Eds.], 2012).

One associated consequence of the choice to use Deane's model is that reading comprehension and writing skills are similarly partitioned into main aspects and sub-aspects and that they allow diagnostic conclusions concerning the various "layers of language usage" distinguished by Deane. Listening and, probably, speaking can be incorporated into the same system (see the Assessment Specification, College voor Toesten en Examen, 2014; Cito, 2013b). The interim objectives drawn-up by the Netherlands Institute for Curriculum Development (SLO) were also used in the construction of the student model for writing skills (2012). These interim objectives correspond well to Deane's model. The SLO nevertheless does not use any underlying student model mentioning the cognitive activities of students. There are ideas concerning what students in vmbo, havo and vwo should know. The Assessment Specification committee has verified that the main aspects together cover the interim objectives. Conversely, it is not easy to assign all of the individual tasks to individual interim objectives.

Based on Deane's model and after consulting teachers by internet, the Assessment Specification committee confirmed a student model for Dutch writing skills in 2014, consisting of four main aspects, each containing three or four sub-aspects (see Table 4-1). Prior to the final version of the student model that was confirmed, modest experimentation with several variants took place. Among other things, sub-aspects and their possible operationalizations were submitted to Dutch teachers, along with the request to reflect on

their recognizability, meaningfulness and usability in teaching practice. The extent of detail was also a point of discussion. For example, the student model had originally consisted of five main aspects and a total of 23 sub-aspects (see Roelofs & Schouwstra [Eds.], 2012, p. 71). The definitive version of this student model was finalized based on discussions within the Assessment Specification committee for Dutch and the internet consultation of Dutch teachers. Brief descriptions of each main aspect and its associated sub-aspects of the student model are presented below.

Table 4-1. Student model for Dutch writing skills (College voor Toetsen en Examens, 2014)

Main aspect	Sub-aspect
1 Rhetorical skills (objective and audience)	1.1 Estimating the prior knowledge and information needs of readers 1.2 Coordinating the tone to the reader (e.g., formal-informal) 1.3 Determining the writing objective (e.g., informing, persuading, explaining, inviting)
2 Text-structure skills (structure)	2.1 Selecting text elements, taking genre into consideration 2.2 Identifying suitable order of text elements and applying proper arrangement and layout 2.3 Applying coherence between text elements (coherence) 2.4 Presenting a standpoint and providing suitable arguments (only havo/vwo)
3 Linguistic skills (word and sentence level)	3.1 Adopting proper sentence construction 3.2 Adopting a suitable and cohesive writing style 3.3 Demonstrating appropriate and varied word usage
4 Orthographic skills (spelling and punctuation)	4.1 Proper spelling of verbs 4.2 Proper application of other rule-guided spelling 4.3 Adopting proper usage of punctuation marks and capitalization

Rhetorical skills (objective and audience)

This concerns the extent to which students are capable of writing well-aimed texts that are adjusted to the requested audience. The tasks focus on the choices that writers must make. The first sub-aspect includes tasks in which students must consider which information should and should not be included in the text. The second sub-aspect concerns the selection and consistent application of the proper tone, including the difference between formal and informal language usage. Tasks for the third sub-aspect assess the selection and use of an appropriate writing objective (e.g., informing, persuading, explaining, inviting).

Text-structure skills (structure)

This main aspect concerns the extent to which students are capable of constructing their texts into a logical, coherent whole. The tasks focus on choices that writers must make when structuring their texts. The first sub-aspect concerns the selection of proper text elements within the genre of the text. The second sub-aspect includes tasks in which students must apply a logical order to the text elements, in addition to applying the associated layout aspects. For havo and vwo, it also includes the use of sub-headings. The third sub-aspect contains tasks focusing on the application of internal coherence between the text elements (e.g., through content words and word fields). The fourth sub-aspect contains tasks in which students must state a standpoint with appropriate arguments (only for havo and vwo), including tasks in which they must provide arguments for and against a standpoint.

Linguistic skills (word and sentence level)

This aspect concerns the formulation of words and sentences. The first sub-aspect includes tasks in which students must recognize and possibly correct faulty sentence construction. The second sub-aspect concerns the use of the proper words to promote coherence in a given text (e.g., conjunctions and reference words). The tasks for the third sub-aspect assess whether a writer is able to select the proper word within a given context, including tasks with figurative language usage for students in havo/vwo.

Orthographic skills (spelling and punctuation)

This concerns the extent to which students are capable of proper spelling and punctuation. The first sub-aspect contains tasks on the spelling of verbs (e.g., finite verbs, infinitives, the past participles of regular and irregular verbs, and attributive verbs). The second sub-aspect includes tasks on the spelling of rule-guided words, focusing on the spelling issues listed in the referential framework for language (e.g., plural formation, diminutives, and suffixes). Tasks associated with the third sub-aspect assess the proper usage of punctuation marks and capitalization.

As previously noted, activities in the preparatory phase included the development of a student model comprising 23 sub-aspects (Roelofs & Schouwstra, 2012). Further examination revealed that several of the sub-aspects in that variant exhibited so much substantive overlap that the student model could be condensed. The student model for writing skills was ultimately limited to 12 sub-aspects for vmbo and 13 sub-aspects for havo/vwo (College voor Toetsen en Examens, 2014). This choice was also made with regard to the practical consequence that it would be necessary to develop a large number of tasks for each sub-aspect in order to ensure the proper measurement of each sub-skill separately.

4.1.2 Operationalization of Dutch writing skills

To operationalize writing skills, several tasks were constructed for each sub-aspect. This section starts by discussing the course of the construction process, followed by a discussion of the various sub-aspects for writing skills and information on several specific features of the types of tasks used.

The construction process

For the development of the DET for writing skills, 545 tasks (with 1900 responses) were created and tried out, either in one of the two pre-tests or as seed tasks. As described in Chapter 2, the tasks were developed by two construction groups consisting of secondary teachers (one for vmbo and one for havo/vwo), based on the sub-aspects. The tasks that had been developed were discussed with the teachers and adjusted as needed. A round of screening, in which each task was assessed by at least two assessment experts and adjusted as needed, followed this. After the tasks had been entered into the item bank (Questify), each task was submitted to the confirmation committee for the DET Dutch. Prior to the meeting, the members of this committee stated their impressions of each task to the staff members of Cito. In the meeting, the unanimously approved tasks were endorsed without further discussion. The tasks that had been rejected by the majority were also not discussed in detail, although the reasons why these tasks were considered unsuitable were explained. The tasks that had been identified as suitable in principle, but that were not good enough in their present form were discussed in detail, so that they could be improved and addressed again in a subsequent meeting.

Try-outs were conducted in 2013 and 2014, in which several types of tasks were tested. In addition, several open-ended writing assignments were developed, so that the writing skills of students could be charted in a more “natural” manner. The ultimate goal was to compare the assessments of teachers concerning performance on these open-ended writing tasks with those concerning performance on more closed-ended tasks. Whenever possible, the automated assessment of the student texts was included as a third evaluation in the comparison. Open-ended writing tasks, including the associated issue of evaluation, are discussed in more detail in Chapter 6.

Based on the experiences from this try-out, some tasks were developed further, and new tasks were constructed. The total numbers of approved tasks and responses are presented in Table 4-2 for each aspect, by educational stream. In all, 476 tasks (1596 responses) were approved for administration following the adaptive administration in 2017 (after two pre-tests and the first time seeding). On average, there are 29 tasks those together yield 52 responses for each sub-aspect in a given educational stream. Several tasks were administered in more than one educational stream. Of all tasks, 50% (42% of the responses) were directed toward single educational streams, with the other half of the tasks being suitable for two or three educational streams. The last column of Table 4-2 lists only the numbers of unique tasks (without overlap).

Table 4-2. Number of approved tasks and responses following the 2017 adaptive administration, for each sub-aspect of the Student Model for Dutch Writing Skills

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
NED-1.1	9 (54)	8 (43)	9 (44)	6 (37)	3 (17)	22 (123)
NED-1.2	11 (33)	11 (37)	14 (47)	14 (31)	15 (29)	41 (111)
NED-1.3	13 (50)	17 (72)	11 (51)	14 (25)	18 (31)	43 (118)
NED-2.1	15 (40)	13 (33)	15 (32)	16 (24)	16 (32)	51 (110)
NED-2.2	12 (12)	15 (15)	19 (19)	17 (32)	21 (37)	65 (84)
NED-2.3	12 (50)	13 (50)	14 (53)	9 (28)	7 (30)	30 (110)
NED-2.4				16 (43)	15 (24)	24 (54)
NED-3.1	10 (29)	15 (52)	13 (44)	15 (67)	16 (71)	42 (149)
NED-3.2	15 (56)	15 (62)	10 (48)	9 (30)	11 (32)	42 (161)
NED-3.3	12 (46)	12 (43)	13 (40)	12 (61)	13 (41)	39 (150)
NED-4.1	10 (51)	12 (64)	12 (62)	8 (49)	8 (49)	23 (124)
NED-4.2	13 (72)	14 (79)	14 (73)	12 (70)	11 (64)	31 (172)
NED-4.3	11 (57)	8 (47)	8 (55)	8 (42)	6 (30)	23 (130)
Total	143 (550)	153 (597)	152 (568)	156 (539)	160 (487)	476 (1596)

Note: In the table, tasks loading on more than one sub-aspect are mentioned under only one sub-aspect.

Types of tasks

The 13 sub-aspects within the concept of writing skills are assessed using a large number of different types of tasks, including drop-down tasks, dragging tasks, short open-ended tasks, and correction tasks. For each task, consideration was given to the nature of the skill to which the task appealed and the type of interaction that would most closely approximate it. For example, short open-ended tasks fit quite well with tasks calling for spelling, and dragging is suited to tasks focusing on structure and word order.

Certain types of tasks were better suited than others for use with certain sub-aspects, as they appeared to offer a better representation of the mental activities that occur during the writing process. For example, for Sub-aspect 3.3, *Demonstrating appropriate and varied word usage*, we often used types of tasks in which students could select the most suitable word from a scrolling menu (drop-down task), based on the assumption that this type of task offers a good representation of the manner in which, during the writing process, a writer can choose from a variety of words or expressions, all of which are technically suitable, although one is preferable within the specific context. For Sub-aspect 1.1, *Estimating the prior knowledge and information needs of readers*, we used multiple-response tasks more often, as they simulate the situation in which, during the planning phase of the writing process, writers reflect on which pieces of information are relevant to mention in the text, given the writing objective and reading audience. In this way, we carefully selected the types of tasks that seemed best suited to offer an accurate simulation of the natural writing process, as it has been described in theories on writing skills (cf. Flower & Hayes, 1981; Bereiter & Scardamalia, 1987; Deane 2011).

To ensure the best possible approximation of particular cognitive sub-activities, specific new types of tasks were developed at our request (see also Chapter 2), including the paragraph task, in which students must use the mouse to identify the manner in which a given text should be divided into paragraphs. With each click of the mouse, a new paragraph break appears in the indicated location, in the same manner as in an ordinary word processor. To correct mistakes, the student can click in the same location, thereby deleting the new paragraph break. Another type of task that was developed specifically for this project is the correction task. In this type of task, students can use the mouse to indicate errors in formulation. This opens a small input field in which they can enter the correction. This type of task is used primarily for Sub-aspects 3.3, *Demonstrating appropriate and varied word usage*, *4.1, Proper spelling of verbs,* and *4.2, Proper application of other rule-guided spelling.*

There are still wishes regarding the types of tasks available. In particular, the assessment experts and construction groups feel a need for a type of task in which students can demonstrate that they are capable

of detecting particular problems with sentence construction and resolving them in more than one way. For example, this is the case with “incongruity errors,” in which the subject and the finite verb do not refer to the same quantity. This type of error can be corrected by changing the subject from singular to plural (or *vice versa*) or by changing the finite verb from plural to singular (or *vice versa*). Both options result in a correct answer. A similar situation applies to errors resulting from the incorrect use of reference words. Errors with reference words and numerical congruity are very common in student texts (cf. Dirksen, Schellens, & Schuurs, 1987). Within the framework of the current project, it proved overly ambitious to develop a type of task that would be capable of charting this sub-aspect.

The total number of approved tasks is presented in Table 4-3, broken down by type of task. As indicated by these figures, marking tasks, multiple-choice tasks, drop-down tasks, and dragging tasks are the most commonly used types of tasks.

Table 4-3. Number of approved tasks and responses following the 2017 adaptive administration, by type of task

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
Paragraph task	2 (2)	3 (3)	4 (4)	3 (6)	3 (6)	11 (14)
Correction task	6 (30)	7 (34)	9 (46)	19 (121)	21 (121)	34 (191)
Drop-down task	27 (94)	25 (95)	25 (97)	15 (47)	13 (40)	70 (243)
Hotspot task	0 (0)	0 (0)	0 (0)	2 (10)	2 (10)	2 (10)
Short open-ended task	18 (95)	23 (126)	19 (106)	10 (56)	8 (47)	39 (212)
Marking task	23 (74)	28 (86)	28 (86)	37 (111)	35 (102)	91 (261)
Matrix task	17 (107)	18 (108)	13 (78)	9 (65)	6 (34)	38 (241)
Multiple-choice task	15 (15)	17 (17)	18 (18)	37 (37)	43 (43)	88 (88)
Multiple-response task	7 (40)	7 (41)	8 (47)	7 (47)	5 (32)	23 (139)
Dragging task	20 (85)	18 (80)	16 (74)	11 (33)	18 (46)	48 (165)
Ordering task	8 (8)	7 (7)	12 (12)	6 (6)	6 (6)	32 (32)
Total	143 (550)	153 (597)	152 (568)	156 (539)	160 (487)	476 (1596)

Operationalization of Dutch writing skills' aspects

Main Aspect 1 Rhetorical skills (objective and audience)

In the first pre-test, the vmbo tasks for Main Aspect 1 had a good average level of difficulty (average p-value between.57 and.70). Although the tasks for this main aspect were somewhat simpler in the second pre-test, their discriminating power was good. In the second pre-test (in 2016), there was little difference in the level of difficulty between the separate tasks within this main aspect. One factor that probably contributed to this is the agreement that was made after the first pre-test to state the number of answers to be detected in the question for vmbo (like in the task in Figure 4-1). For tasks assessing Main Aspect 1, however, this does not always seem desirable, as it could potentially provide too much direction to the students.

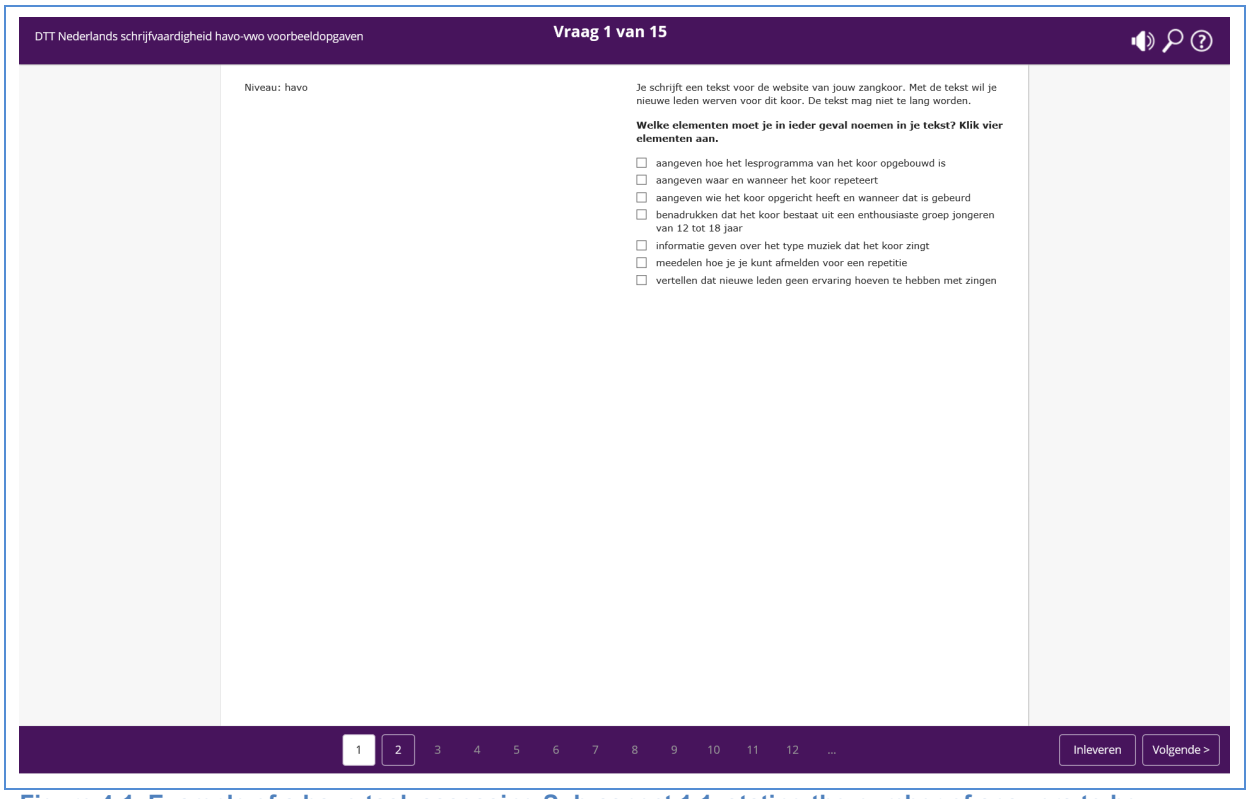


Figure 4-1. Example of a havo task assessing Sub-aspect 1.1, stating the number of answers to be detected. The student has to select *four* pieces of information (out of seven) that should be used in a website text for recruiting new members of a choir.

For students in havo and vwo, the first pre-test (in 2015) revealed that the tasks for Main Aspect 1 were relatively simple (average p >.80). In the construction of the tasks for the second pre-test (in 2016), the attempt was made to construct more difficult tasks for this aspect. This was reasonably successful for Sub-aspects 1.2, *Coordinating the tone to the reader,* and 1.3, *Determining the writing objective*. With regard to the tasks for Sub-aspect 1.1, *Estimating the prior knowledge and information needs of readers*, the tasks were still relatively easy, albeit less so than during the first pre-test. One possible explanation could be that students in the third year of havo and vwo are already quite capable of estimating the prior knowledge of readers and what their information needs are. Another explanation could be that it is difficult to construct tasks for this sub-aspect that are substantively challenging enough for students in havo and vwo, while also generating complete consensus in their evaluation. Because the tasks are closed-ended due to the automated scoring, it is always necessary to select unambiguous tasks, the answers to which are indisputably correct and the distractors are indisputably wrong. For this reason, the tasks for students in havo/vwo on this sub-aspect are likely to be relatively simple.

Main Aspect 2 Text-structure skills (structure)

The level of difficulty and the discriminating power of the tasks assessing Sub-aspects 2.1, *Selecting text elements, taking genre into consideration,* and 2.3, A*pplying coherence between text elements,* are good. In contrast, the tasks for Sub-aspect 2.2, *Identifying suitable order of text elements and applying proper arrangement and layout,* proved relatively difficult. Dragging a number of paragraphs in a given order is clearly a difficult task, particularly for students in vmbo. There is evidence that the “freezing” of a number of paragraphs – which provides more direction to the student – and reducing the number of paragraphs to be dragged affect the level of difficulty. Tasks for which this was the case proved easier than those for which this was not the case (see Figure 4-2 for an example of a task assessing Sub-aspect 2.2 in which two paragraphs are “frozen”). The genre to which the text belongs also appears to influence the level of

difficulty. Flanking research on this sub-topic has been initiated, and a report will be published in March 2018 (Schreurs & Schuurs, forthcoming).

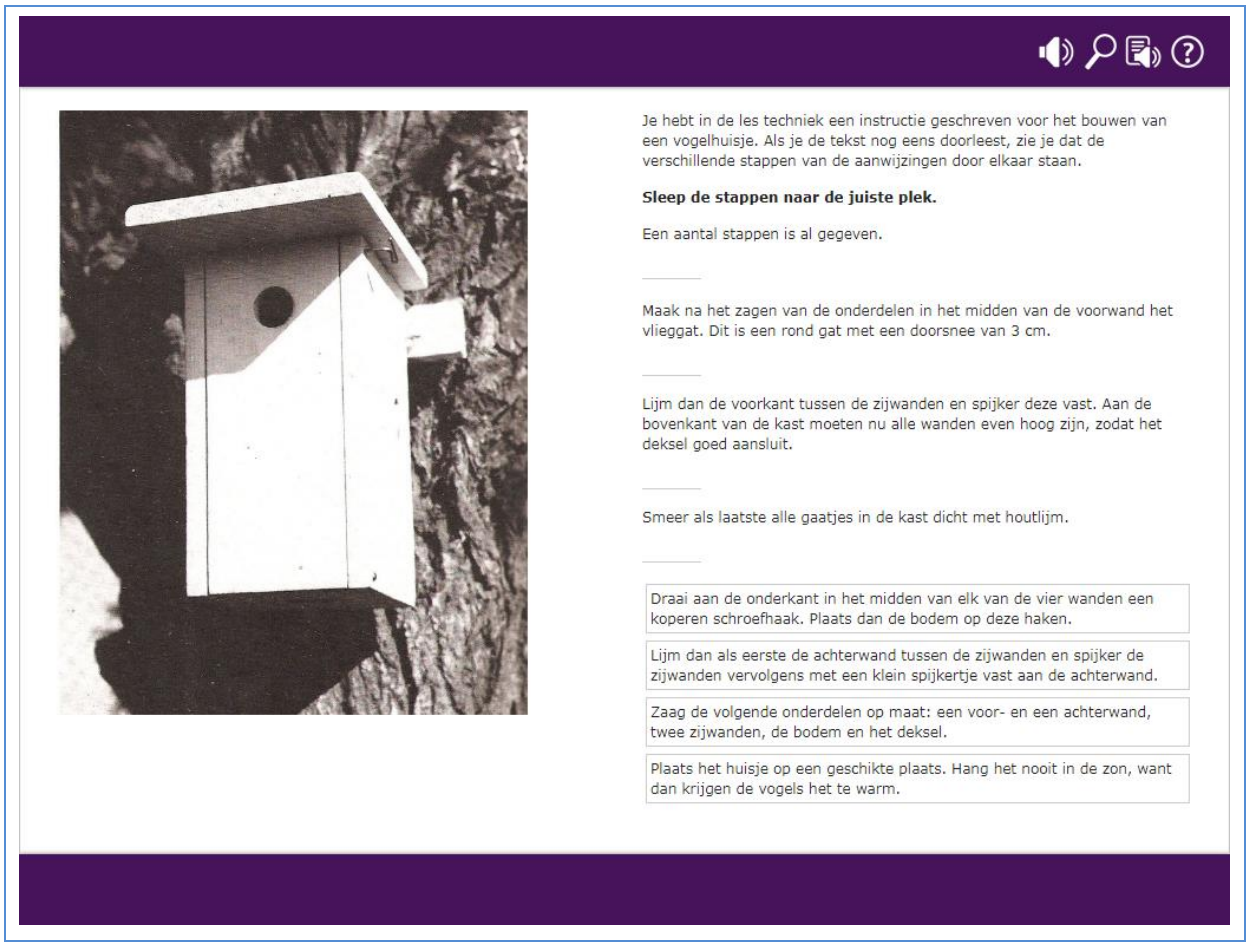


Figure 4-2. Example of a vmbo-kb task assessing Sub-aspect 2.2 in which three paragraphs have been "frozen". In an instruction for building a bird house three steps have been placed in the correct place. Students have to drag the other steps in the correct place of the instruction.

For havo/vwo, the tasks for Sub-aspect 2.2, *Identifying suitable order of text elements and applying proper arrangement and layout,* and Sub-aspect 2.3, *Applying coherence between text elements,* proved much more difficult in 2016 than they had been in 2015. No conclusions can be drawn from this finding, however, given that only a limited number of tasks were administered for Sub-aspect 2.2 in de first pre-test, while the numbers for Sub-aspect 2.3 were quite low in 2016.

In the results from the 2015 pre-test for havo/vwo, it is interesting to note that, contrary to expectations, a relatively large number of tasks were answered correctly for Sub-aspect 2.4, *Presenting a standpoint and providing suitable arguments* (tested only for havo/vwo). This was due primarily to simple distractors in the multiple-choice and multiple-response tasks. In accordance with the recommendations of the Assessment Specification committee, the first pre-test consisted exclusively of tasks in which students had to be able to identify the standpoint in an argument. Because the confirmation committee deemed this assignment too simple for havo/vwo, the arsenal of tasks was expanded. In deviation from what is included in the interim objectives (SLO, 2012), but at the explicit request of teachers and in dialogue with the DET confirmation committee, the 2016 pre-test surveyed other aspects as well, including

- Selecting the proper standpoints and arguments in a short text, using drop-down or other types of tasks;
- Completing argumentation schemes, using dragging tasks in which the central standpoint and the associated arguments had to be dragged to the proper location in an argumentation scheme;
- Identifying which arguments do or do not fit with a given standpoint, using matrix tasks.

Although the number of tasks administered in the 2016 pre-test was limited, it seems reasonable to conclude that drop-down and dragging tasks are particularly good alternatives for administration to students in havo and vwo (see Figure 4-3 for an example of a task assessing argumentative skills).

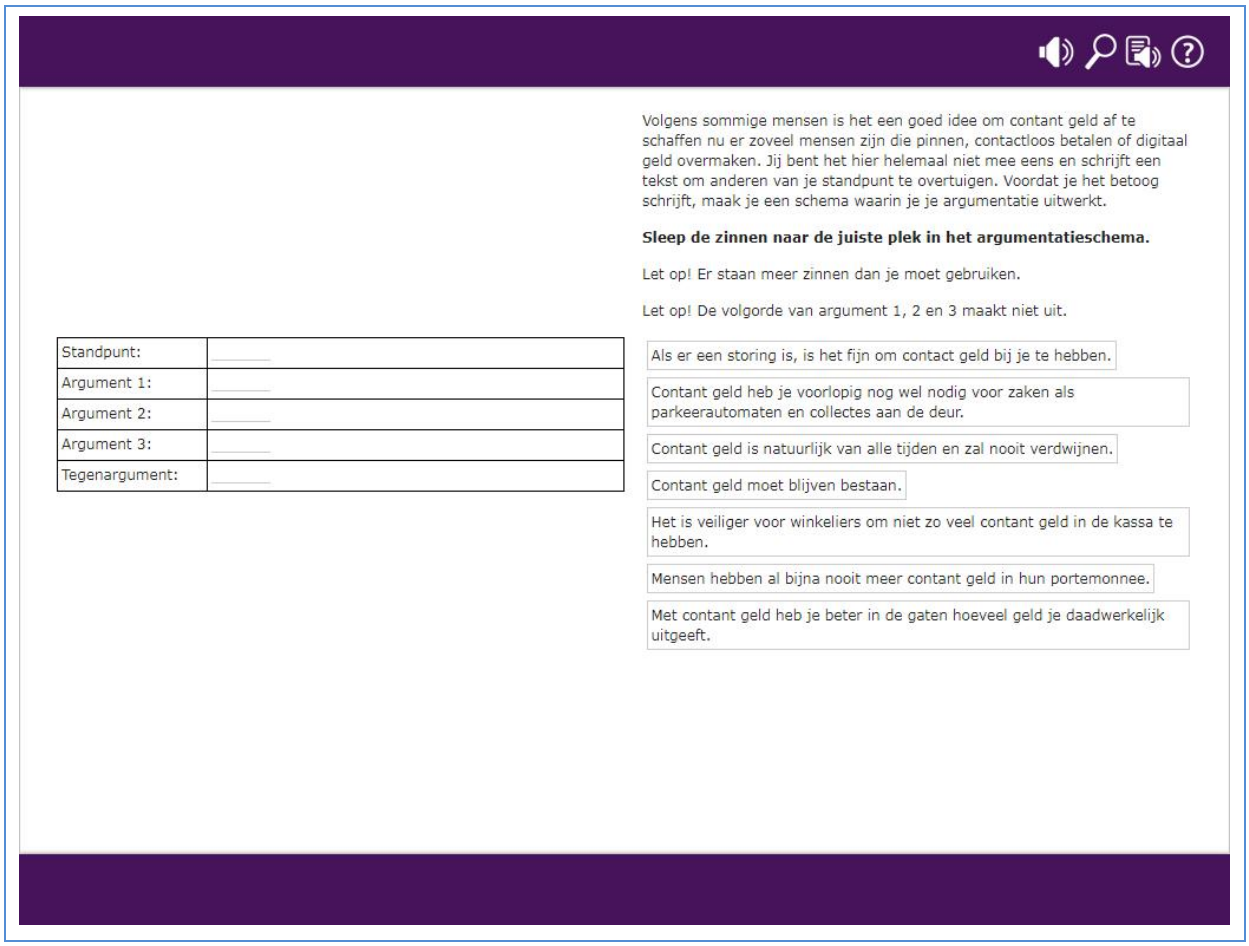


Figure 4-3. Example of an argumentation scheme assessing Sub-aspect 2.4 for havo. Out of seven sentences students have to select a standpoint, three agruments and a counter argument regarding contactless paying.

Main Aspect 3 Linguistic skills (word and sentence level)

For Main Aspect 3, the average level of difficulty of the tasks is good for all educational streams. The average discriminating power is also good. In 2015, tasks for Sub-aspect 3.1, *Adopting proper sentence construction,* were relatively difficult. For vmbo, this was due in part to the use of a number of tasks for all educational streams, which proved too difficult for vmbo-bb. In the 2016 pre-test, the tasks for Sub-aspect 3.1 were still somewhat more difficult than those for the other sub-aspects, although the difference was no longer so great. In addition, the identification of mistakes like numerical incongruity between subject and finite verb or incorrect/unclear references proved relatively difficult for vmbo students. In contrast to the case with Main Aspect 1, the agreement to state the number of elements to be detected appears to have been the right choice here: it provides somewhat more support to students, without producing tasks that are too easy (see Figure 4-4). The selected forms of questioning and types of tasks appear to be a good way to assess this sub-aspect.

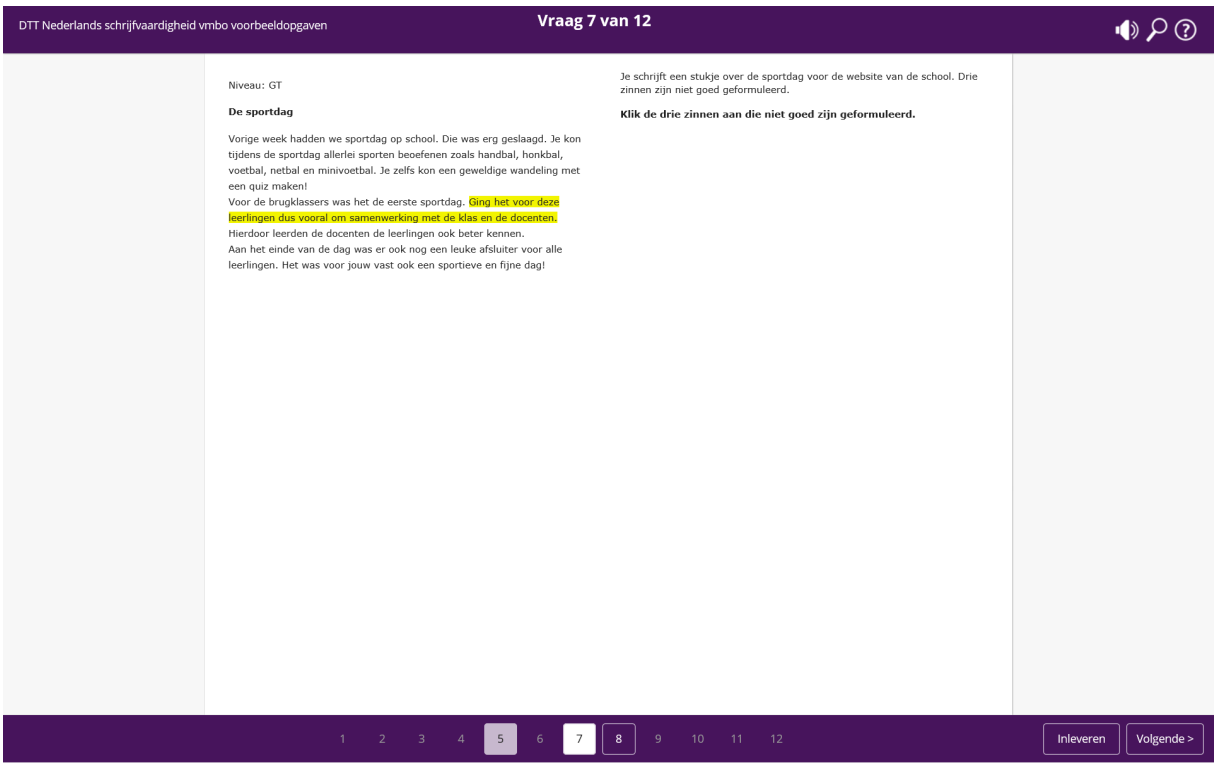


Figure 4-4. Example of a vmbo-kb GT task assessing Sub-aspect 3.1, stating the number of elements to be detected in the question. In the text, students have to select *three* sentences that are formulated incorrectly.

It is interesting to note that students in havo and vwo had considerable difficulty with tasks for Sub-aspect 3.1, *Adopting proper sentence construction* during the pre-test in 2015. In these educational streams, this sub-aspect was assessed primarily using tasks in which students had to detect and correct errors in sentence construction and using tasks in which several sub-aspects are tested at the same time. Both of these tasks proved difficult for students. The most obvious explanation for this finding is that most types of sentence-construction errors that were operationalized in tasks were not treated until the higher years of secondary education, as revealed in an analysis of the three most commonly used teaching methods for Dutch during the preliminary study (Roelofs & Schouwstra, 2012). In contrast, the 2016 pre-test primarily involved the administration of tasks in which students were asked only to detect errors. Students are apparently better able to work with this type of task. Error recognition is nevertheless only the first step. In the context of writing skills, it is important to be able to correct errors and, ultimately, to avoid making them (cf. Schuurs, 1990). With regard to sentence construction, students are apparently still developing in the higher years of secondary education.

In the 2016 pre-test, the tasks for Sub-aspect 3.2, *Adopting a suitable and cohesive writing style,* also appeared to be somewhat easy. Given the limited number of tasks that were administered in that pre-test, however, few conclusions can be drawn from this finding.

Main Aspect 4 Orthographic skills (spelling and punctuation)

For this main aspect as well, the average level of difficulty and discriminating power of the tasks were good. For havo-vwo the discriminating power of these tasks was even the best of all aspects.

In the 2015 pre-test, Sub-aspect 4.3 *Adopting proper usage of punctuation marks and capitalization* reflected a high level of difficulty. To assess this skill in the most direct manner, a choice was made to administer short open-ended tasks for the 2015 pre-test, in which students had to insert punctuation marks in a given text themselves. However, this generated a greater variety of answers than had been anticipated. For example, in addition to entering the intended punctuation mark, some students repeated words or inserted missing spaces. The automated evaluation of all of these variants, which were essentially correct,

proved practically impossible. For this reason, other types of tasks were used to assess this sub-aspect in the 2016 pre-test, including tasks in which students had to drag punctuation marks from a “punctuation board” to the proper locations in a text or in which they had to select the right punctuation mark from a drop-down menu. This form appeared to be much better suited for vmbo-gt (average p-values between.66 and.79). Because the drop-down tasks proved somewhat simple for students in havo and vwo, one of the two tasks for vwo was also rejected after the pre-test. Dragging punctuation marks proved to be a somewhat more difficult variant of the question, and it might offer possibilities for tasks for further testing. An example of a task involving punctuation marks is included in Figure 4-5.

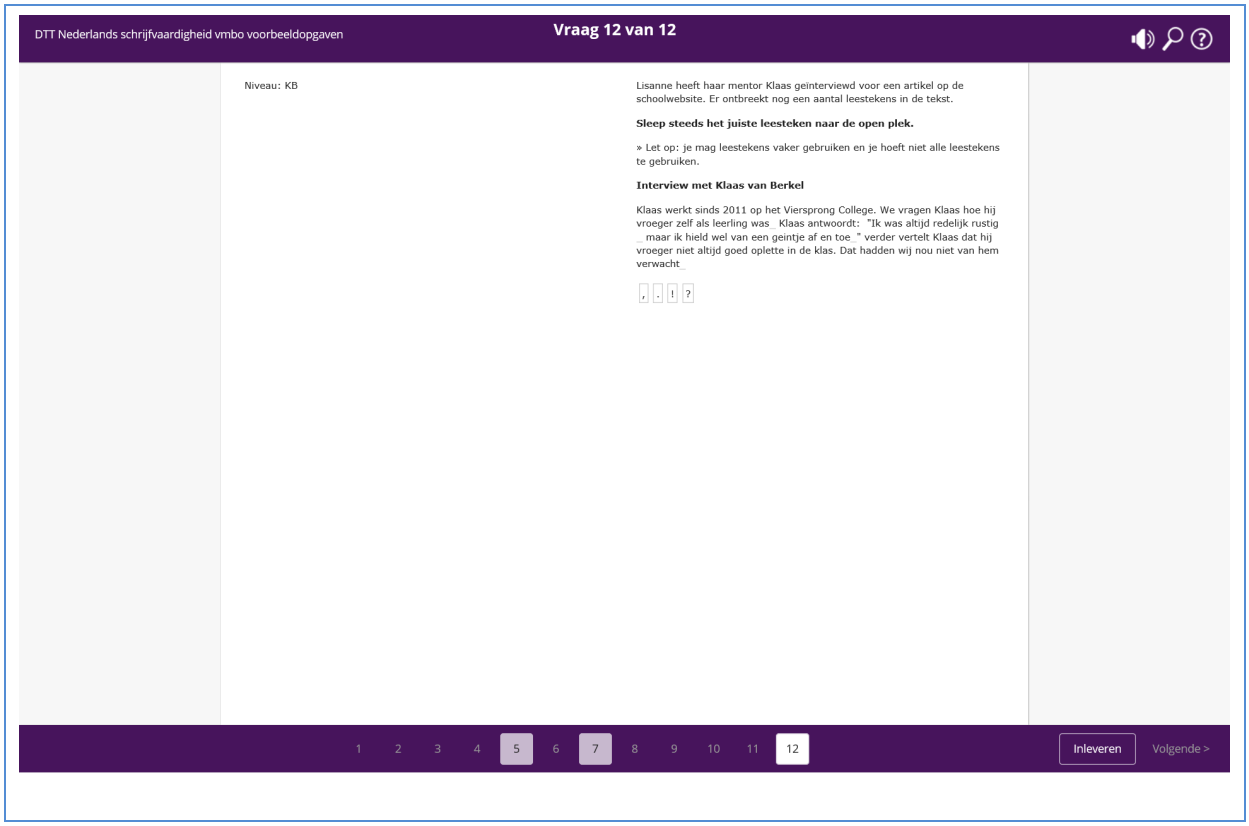


Figure 4-5. Example of a task in which students have to drag punctuation marks to the proper location in the text

In all, 13% of the tested tasks (16% of the responses) were rejected following the two pre-tests and the first time seeding (see Table 4-4 and Table 4-5). It is interesting to note that no tasks were rejected for Sub-aspect 2.3, *Applying coherence between text elements*, and only one was rejected for 2.4, *Presenting a standpoint and providing suitable arguments,* and 4.1, *Proper spelling of verbs* (see Table 4-4). Relatively many tasks were rejected for sub-aspect 1.1, *Estimating the prior knowledge and information needs of readers*, primarily because they were too easy (as described previously). A relatively large number of marking tasks were also rejected. This was a new type of task, with which the construction groups and assessment experts had yet to gain experience.

Table 4-4. Number of rejected tasks and responses following the 2017 adaptive administration, by sub-aspect of the student model

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
NED-1.1	3 (13)	5 (30)	5 (23)	7 (48)	7 (48)	15 (91)
NED-1.2	6 (24)	3 (10)	2 (5)	3 (7)	2 (6)	9 (31)
NED-1.3	1 (3)	1 (3)	1 (3)	1 (7)	1 (7)	2 (10)
NED-2.1	4 (6)	3 (3)	3 (3)	2 (4)	1 (1)	6 (10)
NED-2.2	5 (7)	3 (3)	3 (3)	0 (0)	0 (0)	9 (11)
NED-2.3	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
NED-2.4				1 (6)	0 (0)	1 (6)
NED-3.1	2 (6)	2 (6)	2 (7)	2 (4)	2 (4)	5 (13)
NED-3.2	1 (1)	2 (6)	3 (10)	1 (1)	0 (0)	5 (16)
NED-3.3	0 (0)	0 (0)	1 (6)	2 (16)	1 (8)	3 (22)
NED-4.1	1 (4)	1 (4)	0 (0)	0 (0)	0 (0)	1 (4)
NED-4.2	2 (4)	2 (4)	2 (4)	2 (8)	2 (8)	4 (12)
NED-4.3	4 (36)	5 (47)	6 (59)	2 (19)	2 (19)	9 (81)
Total	29 (104)	27 (116)	28 (123)	23 (120)	18 (101)	69 (307)

Note. In the table, tasks loading on more than one sub-aspect are mentioned under only one sub-aspect.

Table 4-5. Number of rejected tasks and responses following the 2017 adaptive administration, by type of task for the intended educational streams

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
Paragraph task	3 (5)	2 (2)	2 (2)	0 (0)	0 (0)	4 (6)
Correction task	4 (36)	4 (36)	4 (43)	3 (26)	3 (26)	8 (72)
Drop-down task	3 (10)	1 (2)	1 (6)	3 (9)	2 (6)	7 (25)
Short open-ended task	1 (4)	1 (4)	0 (0)	0 (0)	0 (0)	1 (4)
Marking task	9 (16)	10 (29)	11 (32)	7 (16)	6 (15)	20 (56)
Matrix task	3 (19)	4 (25)	2 (10)	3 (27)	3 (27)	9 (64)
Multiple-choice task	0 (0)	0 (0)	1 (1)	2 (2)	1 (1)	3 (3)
Multiple-response task	1 (4)	1 (9)	2 (17)	5 (40)	3 (26)	8 (61)
Dragging task	2 (7)	2 (7)	3 (10)	0 (0)	0 (0)	3 (10)
Ordering task	3 (3)	2 (2)	2 (2)	0 (0)	0 (0)	6 (6)
Total	29 (104)	27 116)	28 (123)	23 120)	18 (101)	69 (307)

4.2 Dutch reading comprehension

4.2.1 Student model for Dutch reading comprehension

Analogous to the model for writing skills, a student model was formulated for reading comprehension. As was the case for writing skills, the initial proposal contained five main aspects. With an eye to the necessity of limiting the number of sub-aspects to be assessed – it had to be possible to construct and pre-test a sufficient number of tasks for each identified sub-aspect – the original model was condensed to a student model with four main aspects and a total of 11 sub-aspects (see Table 4-6). After a consultation of the

educational sector by internet, the Assessment Specification committee confirmed this student model in January 2014.

Table 4-6. Student model for Dutch reading comprehension (College voor Toetsen en Examens, 2014)

Main aspect	Sub-aspect
1 Rhetorical skills (objective and audience)	1.1 Recognizing and interpreting the nature of the source of information, the general writing objective and the intention of specific text fragments 1.2 Interpreting information from texts and drawing conclusions 1.3 Assessing information (and sources) for reliability and usability for the reading objective and substantive quality
2 Text-structure skills (structure)	2.1 Distinguishing main issues from side issues 2.2 Recognizing the construction and coherence of text elements 2.3 Recognizing and understanding essential text elements
3 Linguistic skills (word and sentence level)	3.1 Recognizing and understanding coherence between sentences 3.2 Understanding the meaning of complex text fragments, formulations and words that are relevant to the text[6]
4 Strategic skills (reading strategies)	4.1 Systematically looking up information that is useful, given the reading objective 4.2 Using external features of a text to understand the structure and reasoning 4.3 Deriving the meaning of unfamiliar words based on fragments of the word, sentence or text

In this student model, several clarifications and specifications were added by the Assessment Specification committee. Sub-aspect 1.1 can refer to the following writing objectives: informing, instructing or arguing. Sub-aspect 2.2 refers to the coherence or *substantive* consistency of a text, and 3.1 has to do with the cohesion (i.e., morphosyntactic consistency) of the text. Cohesion can be strengthened through the use of reference words, and lexical consistency can be strengthened by synonyms, hyponyms, and conjunctions. Professional writers usually use cohesion deliberately in order to highlight substantive coherence. For Sub-aspect 2.3, “essential text elements” refer to facts and opinions; standpoints and arguments; and schemes, tables, and charts. Sub-aspect 4.1 concerns looking up information in the media center, on the internet, and in reference works, and 4.2 involves using such matters as title, sub-headings, and the use of various fonts and blank spaces to understand the structure and reasoning.

The partial parallel with the student model for writing skills is remarkable. As with Writing, the first three main aspects concern skills relating to rhetoric, text structure, and linguistics. This student model includes a main aspect related to reading strategies instead of orthographic skills – which, in the case of Reading, do not appear to be testable at the secondary education level.

4.2.2 Operationalization of Dutch reading comprehension

The main and sub-aspects of the student model are elaborated as follows in the DET. The Public Version of the Assessment Specification for the DET (College voor Toesten en Examens, 2014) presents several examples of tasks. This section is therefore limited to providing additional details concerning the sub-skills that we intend to measure, along with comments concerning the manner in which the sub-aspects are operationalized.

Main Aspect 1 Rhetorical skills (objective and audience)

The main aspect *Rhetorical skills (objective and audience)* concerns the extent to which students are capable of recognizing the objective and audience for which a text has been written, which information is being provided, and the extent to which the information (and sources) used in the text can be assessed.

[6] Including the understanding of figurative language usage (only for havo/vwo)

The following types of tasks have been incorporated into the category of *Rhetorical skills*:

- Tasks focusing on the recognition and interpretation of the nature of the information source, the general writing objective, and the intentions of specific text fragments: does the reader recognize the type of text and the general writing objective (e.g., informing, instructing, or arguing), and can the reader indicate the intentions of particular text fragments?
- Tasks in which information from the text must be interpreted and conclusions must be drawn: can the reader interpret specific information and opinions from the text, and does the reader recognize conclusions contained in the text (havo) or can the reader draw independent conclusions based on information from the text (vwo)?
- Tasks focusing on the assessment of information (or sources) for reliability, usability, and substantive quality: can the reader indicate whether the information (or source) is reliable or usable with regard to the reading objective, and can the reader indicate whether the quality of the information (or sources) is sufficient or insufficient?

Main Aspect 2 Text-structure skills (structure)

The main aspect *Text-structure skills (structure)* concerns the extent to which students can explain the structure of a text and use it to understand and interpret the text.

The following types of tasks have been incorporated into the category of *Text-structure skills*:

- Tasks focusing on distinguishing main aspects from side issues: can the reader identify the most important information in a text?
- Tasks focusing on the construction of and consistency between text elements: can the reader recognize textual connections and substantive connections?
- Tasks focusing on the recognition and comprehension of essential text elements: can the reader recognize and understand facts, opinions, standpoints, as well as arguments, schemes, tables, and charts?

Main Aspect 3 Linguistic skills (word and sentence level)

The main aspect *Linguistic skills (word and sentence level)* concerns the extent to which students understand sentences and words.

The following types of tasks have been incorporated into the category of *Linguistic skills*:

- Tasks focusing on the recognition and comprehension of associations between sentences: can the reader use reference words, conjunctions, synonyms, and hyponyms to assign meaning to sentences and their relationships to each other?
- Tasks focusing on complex text fragments, formulations, and words that are relevant to the text: does the reader understand the meaning of complex words and sentences? Does the reader understand the use of figurative language (only for havo and vwo)?

Main Aspect 4 Strategic skills (reading strategies)

The main aspect *Strategic skills (reading strategies)* concerns the extent to which students can use strategies in the process of assigning meaning.

The following types of tasks have been incorporated into the category of *Strategic skills (reading strategies)*:

- Tasks focusing on systematically looking up usable information: can the reader select the right information (e.g., in the media center, on the internet, and in reference works) with an eye to the reading objective?
- Tasks focusing on the use of external features of a text to understand the structure and reasoning: can the reader use the title, sub-headings, the use of various fonts, blank spaces, illustrations, and other elements to explain the structure of the text and/or to understand the reasoning of the text?
- Tasks focusing on deriving the meaning of unfamiliar words: can the reader derive the meaning of unfamiliar words from parts of the word, sentence, or text?

During the discussions of the Assessment Specification committee and during the task-construction process, the distinction between Sub-aspects 1.2, 2.3, 3.2, and 4.3 proved difficult to operationalize. For this reason, the differences between these sub-aspects are explained here in detail.

- Sub-aspect 1.2 concerns the interpretation of the text as a whole. In essence, tasks for this sub-aspect amount to the question, “What is the text actually saying?”
- Sub-aspect 2.3 focuses on the specific understanding of a number of specifically mentioned text elements (the distinction between fact and opinion, between arguments and conclusion, and between the understanding of schemes, tables, and charts).
- Sub-aspect 3.2 concerns the understanding of a truly complex text fragment (e.g., a truly difficult sentence or difficult words that are really relevant to the text).
- In contrast, Sub-aspect 4.3 concerns unfamiliar words in the general sense: students must be capable of deriving the meaning of these words from parts of the word, sentence, or text, although the words need not play an important role in the text. One distinguishing feature of Sub-aspect 4.3 is thus that the student must be able to distill the meaning of the word from the context.

The situation in which several prototypical sample tasks for writing skills could be suitable as sample tasks for reading comprehension had already been addressed in the Assessment Specification committee prior to confirmation of the student model for Reading. In response, the Assessment Specification committee elaborated a proposal in which Reading and Writing were combined into the single common domain of Written Language Skills, which would combine a large number of sub-aspects as currently formulated separately for Reading and for Writing. Justification for doing so could be found from the theoretical perspective (cf. e.g., Shanahan, 2016), as well as from teaching practice. This globalizing approach was eventually abandoned, as it is in direct opposition to the objective of making diagnostic assessment as specific as possible. Consequently, throughout the entire course of the project, the operationalizations for reading comprehension and for writing skills were very close to each other in some cases.

In all, 506 tasks were developed (see Table 4-7) for Reading Comprehension. More than half of the tasks are multiple-choice, and more than one fourth are marking tasks (see Table 4-8). For a description of the types of tasks used for reading comprehension, we refer to the examples in the previous section and in Chapter2.

Table 4-7. Number of confirmed tasks and responses for each sub-aspect of the Student Model for Dutch Reading Comprehension

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
NED 1.1	12 (13)	12 (21)	14 (17)	16 (24)	15 (23)	54 (78)
NED 1.2	16 (36)	19 (39)	15 (28)	29 (51)	24 (31)	81 (140)
NED 1.3	5 (5)	11 (13)	3 (3)	11 (31)	10 (15)	31 (53)
NED 2.1	6 (10)	12 (19)	9 (11)	14 (28)	12 (19)	40 (66)
NED 2.2	2 (2)	5 (7)	2 (2)	16 (16)	19 (20)	32 (35)
NED 2.3	4 (5)	17 (32)	16 (36)	13 (16)	10 (16)	41 (70)
NED 3.1	10 (13)	10 (14)	3 (4)	9 (9)	11 (13)	31 (37)
NED 3.2	8 (8)	14 (14)	14 (17)	23 (24)	23 (24)	57 (61)
NED 4.1	23 (29)	14 (16)	3 (4)	20 (23)	15 (17)	59 (70)
NED 4.2	5 (5)	5 (5)	6 (6)	16 (24)	6 (12)	29 (41)
NED 4.3	7 (8)	15 (16)	19 (19)	17 (17)	11 (11)	51 (52)
Total	98 (134)	134 (196)	104 (147)	184 (263)	156 (201)	506 (703)

Table 4-8. Number of confirmed tasks and responses for Dutch Reading Comprehension, by type of task

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
Categorization task	1 (5)	1 (5)	0 (0)	0 (0)	0 (0)	1 (5)
Drop-down task	0 (0)	1 (1)	0 (0)	1 (2)	0 (0)	2 (3)
Hotspot task	11 (13)	6 (8)	1 (2)	18 (30)	5 (7)	32 (47)
Short open-ended task	0 (0)	0 (0)	0 (0)	1 (1)	2 (5)	2 (5)
Marking task	26 (28)	44 (54)	31 (37)	37 (44)	42 (52)	138 (163)
Matrix task	6 (22)	4 (18)	5 (21)	10 (43)	3 (11)	22 (89)
Multiple-choice task	46 (46)	66 (66)	62 (62)	105 (105)	90 (90)	273 (273)
Multiple-response task	3 (10)	8 (34)	5 (25)	7 (27)	6 (16)	22 (85)
Dragging task	5 (10)	4 (10)	0 (0)	4 (8)	7 (17)	13 (30)
Drag task with image	0 (0)	0 (0)	0 (0)	1 (3)	1 (3)	1 (3)
Total	98 (134)	134 (196)	104 (147)	184 (263)	156 (201)	506 (703)

4.3 Reflection

In retrospect, we can state that the process of developing the DET Dutch instigated a number of important developments. First, it is of importance to note that Dutch Writing skill was divided into 13 sub-aspects in a scientifically justified manner, which many Dutch teachers recognize as important within their subject area. The same applies to the division of the Dutch Reading Comprehension in 12 sub-aspects, in which teachers usually adopt only three levels (macro level, meso level, and micro level), roughly corresponding to the textual level, paragraph level, and sentence level). The categorizations that were created gave teachers an opportunity to provide individual feedback to students on each of the distinct sub-aspects, regardless of the

underlying didactic philosophy and independent of the methods used. Formative evaluation does not end with the assessment tools that have been provided. It begins in the classroom (cf. Wiliam, 2012), and teachers stand to benefit greatly from the views developed in the DET with regard to Writing Skills and Reading Comprehension.

In addition, hundreds of tasks were constructed, which are operationalizations of the sub-aspects that have been distinguished. These tasks provide evidence that the original categorizations, which were based on theoretic insights, are justified. They can apparently be translated into tasks that teachers regard as making sense. In addition, the results of the pre-tests indicate that they have been coordinated at the proper level. The tasks are also of good quality. In the adaptive administration, a good overview of the skills of students can be obtained with relatively few tasks. This is because the discriminating power of the tasks is clearly good, as evidenced in the simulation studies (see Chapter 15).

Several questions remain, however, suggesting the need for further research. We refer to two issues.

In the project, many types of tasks were used, and new types of tasks were developed at the request of the developers. Interviews with schools (College van Toetsen en Examens, forthcoming) revealed that the creative manner of questioning was greatly appreciated. The teachers were enthusiastic about the new forms of tasks and, according to the teachers, the students perceived the tasks as motivating. At the same time, one could wonder whether the diversity of tasks might have had an unintended negative influence on student performance levels. It is conceivable that the intended skills could be measured much more accurately if the student's attention is not continuously disrupted between substantive considerations and the technical issue of how the answer must be provided in yet another new type of task. Research has yet to address these issues, and it would seem worthwhile to use studies based on a thinking-out-loud protocol to determine whether students might be able to work in a more targeted manner in another form of test (e.g., with only three types of tasks).

The relationship between the diagnoses for sub-aspects and the ability to write well is not clear. Although it is tempting to proceed from the assumption that "above level" diagnoses on the sub-aspects indicate high levels of Writing Skills or Reading Comprehension, we cannot be certain that this is the case. Research is needed on the relationship between the diagnoses on sub-aspects in the DET and the overall assessments of teachers concerning the reading and writing performance of their students.

4.4 References

Bereiter, C., & Scardamalia, M. (1987). *The Psychology of Written Composition.* Hillsdale, NJ: Erlbaum.

Cito (2013b). *Diagnostische tussentijdse toets: Voorstudies en evaluaties 2013 [Diagnostic Educational Test: Preliminary studies and evaluations 2013].* Arnhem: Cito.

College voor Toetsen en Examens (2014), *Publieksversie Toetswijzer Diagnostische Tussentijdse Toets voor Nederlands, Engels en Wiskunde [Public Version of the Assessment Specification for the Diagnostic Educational Test for Dutch, English and Mathematics].* Utrecht: College voor Toetsen en Examens. Retrieved from https://www.pilotdtt.nl/documenten/publicaties/2014/12/15/toetswijzer-dtt

College voor Toetsen en Examens (forthcoming). *Eindrapport project DTT [Final report project DTT].* Utrecht: College voor Toetsen en Examens.

Deane, P. (2011). *Writing Assessment and Cognition*. Princeton, New Jersey: Educational Testing Service.

Deane, P., Odendahl, N., Quinlan, Th., Fowles, M., Welsh, C. & Bivens-Tatum, J. (2008). *Cognitive Models of Writing: Writing Proficiency as a Complex Integrated Skill*. Research Report ETS RR-08-55. Princeton, New Jersey: Educational Testing Service.

Dirksen, A., P. J., Schellens, & U. Schuurs (1987). Vragen fouten in de zinsbouw om grammatica-onderwijs? [Do sentence-construction errors call for grammar education?] *Spektator* 16, pp. 380-393.

Flower, L. & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*, 32, 365-387.

Rijlaarsdam, G.C.W., van den Bergh, H., Couzijn, M., Janssen, T., Braaksma, M., Tillema, M., Van Steendam, E. & Raedts, M. (2011). Writing. In S Graham, A Bus, S Major & L Swanson (Eds.), *Application of Educational Psychology to Learning and Teaching*. APA Handbook Volume 3

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

Schreurs, Z., & Schuurs, U. (forthcoming). *Bevraging van tekststructuur: het effect van genre en het aantal te plaatsen elementen op de moeilijkheidsgraad [Surveying text structure: The effect of genre and number of elements to be placed on the level of difficulty].* Arnhem: Cito.

Schuurs, U. (1990). *Leren schrijven voor lezers [Writing for readers].* Doctoral dissertation, University of Twente.

Shanahan, T. (2016). Relationships between Reading and Writing Development. In Ch. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of Writing Research*. New York, The Guilford Press, 2016.

SLO: Nationaal expertisecentrum leerplanontwikkeling [Netherlands Institute for Curriculum Development] (April 2012). *Concept-tussendoelen kernvakken onderbouw vo [Draft interim objectives for core subjects in the lower years of secondary education].* Retrieved from http://www.slo.nl/nieuws/00291/.

Sluijsmans, D., Joosten-ten Brinke, D., & Van der Vleuten, C. (2013). *Toetsen met leerwaarde: een reviewstudie naar de effectieve kenmerken van formatief toetsen [Testing with educational value: A review study on the effective properties of formative testing].* Maastricht: Maastricht University

Wiliam, D. (2012). *Embedded Formative Assessment*. Bloomington: Solution Tree Press.

5 DET English

Ingrid Williams and Wilma Vrijs

5.1 English writing skills

5.1.1 Student model for English writing skills

As is the case with the DET Dutch, the model developed by Deane served as the base for a general student model for language skills (Deane et al., 2008; Deane, 2011). This model cannot be applied to English, however, as it was developed for first-language acquisition. In addition, it does not address several aspects of writing at the level of lower grades. During the preliminary study (Roelofs & Schouwstra [eds.], 2012), five main aspects were assigned to the student model for writing skills: *Tuning to audience and objective*; *Coherence*; *Vocabulary and word usage*; *Grammar, spelling, and punctuation* and *Supporting skills*. In the process of finalizing the student model, the Assessment Specification committee chose to eliminate the fifth aspect (*Supporting skills*), as these skills are implicitly addressed in other sub-aspects.[7] The final student model for English writing skills therefore distinguishes the following four main aspects: *Tuning to audience and objective*; *Coherence*; *Vocabulary and word usage; Grammar, spelling, and punctuation*.

In the interest of identifying sub-aspects, draft interim objectives formulated by the Netherlands Institute of Curriculum Development (SLO) based on the EFR were analyzed for testability. The analysis also drew on research investigating the link between the Modern Foreign Languages curriculum and the European Framework of Reference (Van Hest, Beltman & Kleintjes, 2001). To develop the most complete image possible of customary aspects of English writing skills, we also considered the EFR "can-do statements" and the extent to which the student model for Dutch corresponds to the aspects presented by Meijerink (2008). This ultimately led to the distinction of eight sub-aspects in the student model for English writing skills.

Taken together, the main aspects and sub-aspects constitute the student model for English writing skills, as summarized in Table 5-1. Given that the process of language acquisition occurs in fits and starts, this table should not be interpreted as a sequence or hierarchy. There is no linear structure in skills.

Table 5-1. Student model for English writing skills (College voor Toetsen en Examens, 2014)

Main aspect	Sub-aspect
1. Tuning to audience and objective	1.1 Is capable of tuning tone and register to the audience and writing objective. 1.2 Is capable of using conventions associated with a type of text.
2. Coherence	2.1 Is capable of applying text structure and relationships. 2.2 Is capable of using structure words suited to the task: use of conjunctions and reference words.
3. Vocabulary and word usage	3.1 Is capable of using words and combinations of words suited to the task. 3.2 Is capable of functional variation in word usage (including as a compensating strategy).
4. Grammar, spelling, and punctuation	4.1 Is capable of functional use of word order and sentence construction. 4.2 Is capable of using correct spelling and punctuation.

The main aspect *Tuning to audience and objective* concerns the extent to which students are capable of tuning their texts to the audience and objective in such a way that the texts are communicatively effective: is the message conveyed in the intended manner? The first sub-aspect calls upon the ability to tuning the tone and register to the intended audience and writing objective. The second sub-aspect focuses on the

[7] In sub-aspects 3.2 "Is capable of functional variation in word usage" and 4.1 "Is capable of functional use of word order and sentence construction."

conventions associated with specific types of texts: is the writer capable of opening and closing a letter in the customary manner?

The second main aspect, *Coherence,* concerns the application of logical and coherent structure within a text. The first sub-aspect calls upon the ability to apply text structure and relationships. The second sub-aspect calls upon the ability to use structure words correctly and appropriately. Examples include the use of summaries, references, conjunctions, and reference words.

The third main aspect, *Vocabulary and word usage,* concerns the extent to which students are capable of applying their vocabularies and word usage in order to carry out a task properly: is the writer capable of using words that are suited to the context? The first sub-aspect includes tasks that ask students to use words, combinations of words and idiomatic expressions in various contexts. The second sub-aspect calls upon the ability to achieving functional variation in word usage. This also concerns the use of compensating strategies: is the writer capable of using synonyms and descriptions?

The main aspect *Grammar, spelling, and punctuation* concerns the use of sentence structure, word order, and rules for spelling and punctuation. The first sub-aspect calls upon the ability to apply word order and sentence construction properly. The second sub-aspect focuses on correct spelling and appropriate punctuation.

5.1.2 Operationalization of English writing skills

The tasks were developed by two construction groups consisting of secondary teachers: one construction group for vmbo (preparatory secondary vocational education) and one for havo (senior general secondary education)/vwo (university preparatory education). Tasks were created according to the sub-aspects, with the goal of having each task call for only the specific intended sub-skills. All of the tasks were submitted to a confirmation committee consisting of secondary education experts (see Chapter 2). The pre-tests tested only those tasks that were ultimately approved by the confirmation committee. After the pre-tests, the tasks were once again submitted to the confirmation committee, which either approved or rejected the tasks based on the statistical analyses from the pre-test.

The total numbers of approved tasks and responses are presented in Table 5-2 for each sub-aspect, by educational stream. In all, 731 tasks (1294 responses) were approved for administration following the adaptive administration in 2017 (after two pre-tests and the first time seeding). On average, there are 29 tasks that together yielded 52 responses for each sub-aspect in a given educational stream. Several tasks were administered in more than one educational stream. Of all tasks, 51% (49% of the responses) were directed toward single educational streams, with the other 49% of the tasks being suitable for two or three educational streams. These tasks can be seen in the total number listed in the breakdown by level for each educational stream. They are nevertheless counted only once in the last column for the total number of unique tasks and responses.

Table 5-2. Number of approved tasks and responses following the 2017 adaptive administration, for each sub-aspect of the Student Model for English Writing Skills

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
ENG-1.1	27 (51)	25 (62)	24 (76)	26 (68)	19 (43)	73 (171)
ENG-1.2	22 (48)	26 (56)	23 (47)	29 (76)	19 (44)	75 (178)
ENG-2.1	28 (37)	32 (42)	30 (38)	19 (30)	15 (24)	62 (88)
ENG-2.2	31 (37)	32 (49)	36 (52)	31 (54)	32 (56)	106 (159)
ENG-3.1	31 (71)	30 (80)	27 (68)	46 (86)	40 (74)	113 (235)
ENG-3.2	26 (59)	33 (67)	31 (64)	28 (40)	28 (42)	95 (164)
ENG-4.1	27 (31)	30 (51)	32 (47)	32 (51)	19 (37)	94 (142)
ENG-4.2	30 (34)	38 (54)	33 (47)	35 (53)	29 (42)	113 (157)
Total	222 (368)	246 (461)	236 (439)	246 (458)	201 (362)	731 (1294)

Note: In the table, tasks loading on more than one sub-aspect are mentioned under only one sub-aspect.

The sub-aspects are assessed using a large number of different types of tasks, including the following: drop-down tasks, dragging tasks, short open-ended tasks and tasks in which students must select and re-write text fragments. Because the DET is adaptive, all of the types of tasks developed are suitable for automated assessment. The task types differ within the sub-aspects. Certain types of tasks did prove better suited than others for assessing particular sub-skills. In many cases, the nature of the task determines whether an action feels authentic or whether it is not well suited to a particular requested sub-skill. For each task, consideration was given to the nature of the skill to be called upon and the type of interaction that would lead to the most direct manner of assessing it. For example, short open-ended tasks fit quite well with tasks calling for spelling, and dragging action is highly suited to tasks focusing on structure and word order. Some types of interaction can also make it more difficult to assess particular skills. For example, consider the sub-aspect “Vocabulary and word usage.” This aspect is quite difficult to measure in a short open-ended task without also asking about the spelling of a word. It is impossible to include all spelling variants in the response model. In such cases, the choice was made to use a type of task that does not stand in the way of assessing the specific sub-skill.

The total number of approved tasks is presented in Table 5-3, broken down by type of task. As indicated by these figures, drop-down tasks, multiple-choice tasks, dragging tasks, short open-ended tasks, and ordering tasks are the most commonly used types of tasks.

Table 5-3. Number of approved tasks and responses following the 2017 adaptive administration, by type of task

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks	Number of tasks and responses
Paragraph task	1 (3)	1 (3)	1 (3)	0 (0)	0 (0)	2 (6)
Categorization task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Combination task	1 (2)	2 (4)	1 (2)	3 (9)	3 (8)	6 (16)
Correction task	1 (2)	0 (0)	0 (0)	0 (0)	0 (0)	1 (2)
Drop-down task	55 (83)	63 (133)	71 (133)	89 149)	73 (127)	224 (383)
Hotspot task	0 (0)	0 (0)	0 (0)	2 (8)	2 (8)	2 (8)
Short open-ended task	28 (29)	37 (48)	24 (32)	30 (37)	25 (28)	97 (119)
Marking task	6 (6)	6 (10)	2 (2)	3 (8)	4 (12)	12 (24)
Matrix task	3 (18)	6 (31)	9 (45)	13 (77)	9 (57)	25 (147)
Multiple-choice task	62 (62)	61 (61)	55 (56)	57 (57)	50 (50)	195 (196)
Multiple-response task	6 (29)	7 (28)	7 (27)	6 (32)	1 (7)	15 (70)
Dragging task	35 (95)	36 (95)	39 (96)	29 (67)	22 (53)	98 (241)
Dragging task with image	4 (19)	6 (27)	5 (21)	0 (0)	0 (0)	8 (36)
Ordering task	20 (20)	21 (21)	22 (22)	14 (14)	12 (12)	46 (46)
Total	222 (368)	246 (461)	236 (439)	246 (458)	201 (362)	731 (1294)

In all, 6% of the tested tasks (7% of the responses) were rejected following the two pre-tests and the first-time seeding (see Table 5-4 and Table 5-5). In most cases, these rejected tasks were tasks for which too many students had given the right answer, thus indicating that the tasks were unable to discriminate between more and less skilled students. In addition, several tasks were too difficult and therefore incapable of such discrimination. It is interesting to note that, for Sub-aspect 3.2, “Functional variation in word usage,” only one task was rejected (see Table 5-4). All of the other 95 tasks were approved.

Table 5-4. Number of rejected tasks and responses following the 2017 adaptive administration, by sub-aspect of the Student Model

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks	Number of tasks and responses
ENG-1.1	1 (1)	0 (0)	0 (0)	0 (0)	1 (5)	3 (7)
ENG-1.2	2 (12)	5 (10)	4 (9)	1 (1)	1 (1)	9 (24)
ENG-2.1	0 (0)	0 (0)	2 (2)	1 (3)	0 (0)	4 (6)
ENG-2.2	2 (2)	3 (4)	2 (3)	1 (1)	1 (1)	9 (10)
ENG-3.1	2 (2)	2 (7)	3 (14)	1 (3)	2 (2)	11 (31)
ENG-3.2	0 (0)	0 (0)	1 (5)	0 (0)	0 (0)	1 (5)
ENG-4.1	2 (2)	1 (1)	1 (1)	2 (3)	4 (9)	9 (17)
ENG-4.2	0 (0)	1 (1)	1 (1)	0 (0)	1 (4)	6 (10)
Total	9 (19)	12 (23)	14 (35)	6 (11)	10 (22)	52 (110)

Table 5-5. Number of rejected tasks and responses following the 2017 adaptive administration, by type of task for the intended educational stream

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
Paragraph task	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	1 (3)
Correction task	0 (0)	0 (0)	0 (0)	1 (5)	1 (5)	1 (5)
Drop-down task	2 (2)	3 (4)	3 (4)	10 (19)	4 (11)	16 (27)
Hotspot task	0 (0)	1 (3)	0 (0)	0 (0)	0 (0)	1 (3)
Short open-ended task	3 (3)	2 (2)	1 (1)	3 (4)	1 (1)	9 (10)
Marking task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Matrix task	0 (0)	0 (0)	0 (0)	1 (4)	0 (0)	1 (4)
Multiple-choice task	2 (2)	2 (2)	2 (2)	4 (4)	3 (3)	10 (10)
Multiple-response task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Dragging task	0 (0)	2 (2)	5 (19)	2 (2)	0 (0)	7 (21)
Dragging task with image	2 (12)	1 (6)	1 (7)	0 (0)	1 (6)	4 (25)
Ordering task	0 (0)	0 (0)	2 (2)	0 (0)	0 (0)	2 (2)
Total	9 (19)	12 (22)	14 (35)	21 (38)	10 (26)	52 (110)

5.2 Reading comprehension

5.2.1 Student model for English reading comprehension

The cognitive model developed by Deane – which describes and explains the task processes that play a role in language acquisition – also served as the foundation for the student model for reading comprehension. In addition, links were sought with the Common European Framework of Reference for Languages (CEFR) and the interim objectives, which describe levels of language competence in language-task situations of increasing complexity. Wherever possible, the choice was made to correspond to aspects and underlying theoretical bases of the student model for writing skills. This ultimately generated the following four main aspects: *Targeted reading*; *Recognizing text functions*; *Understanding text structure*; *Vocabulary and Word usage*. In Table 5-6 an overview of the student model of Reading Comprehension can be seen.

The first main aspect, *Targeted reading,* concerns the extent to which students are capable of using reading strategies to find information that is needed in order to achieve a given reading objective. The sub-skills that are tested in other main aspects overlap with these reading strategies. The student applies a given reading strategy in order to answer the tasks within a given sub-aspect. The first main aspect consists of three sub-aspects: *skimming*, *scanning,* and *intensive reading*. For example, tasks for *skimming* call upon the ability to derive the main idea of a text. In contrast, the second sub-aspect, *scanning*, is a strategy that can be applied when details must be extracted from a text: the reader searches selectively for relevant details from the text. Finally, *intensive reading* involves the ability to extract finer nuances in meaning from the text (e.g., distinguishing main aspects and side issues).

The second main aspect, *Recognizing text functions: recognizing the objective and type of the text*, involves the extent to which students are capable of recognizing the objective and audience for which a text was written. It is further subdivided into the following aspects: *recognizing the author's objective* and *using text information and drawing conclusions*. *Recognizing the author's objective* concerns whether the reader is capable of determining whether the author is seeking to inform, persuade, instruct, or amuse. Tasks for the

second sub-aspect should call upon the ability to use comparisons, classifications, and illustrations to find information and draw conclusions about a text.

The third main aspect, *Understanding text structure: extracting information from texts*, involves the skill of using structural elements to extract information from a text. Tasks for the first sub-aspect ask the student to identify and/or express the most important idea in a text: *deriving the main idea of the text*. Tasks for the second sub-aspect call upon the ability to distinguish between main aspects and side issues: is the reader capable of distinguishing important information in a text from less important information? Tasks for the third sub-aspect call upon the ability to identify key phrases. Tasks for the fourth sub-aspect ask the reader to search for detailed information in a text: *recognizing relevant details*. The fifth sub-aspect calls upon the ability to understand the structure of the text (e.g., orderings in time, cause-effect relationships). Tasks for the sixth sub-aspect examine the comprehension of references between sentences. Finally, the last sub-aspect calls upon the ability to understand the meaning of signal words.

The fourth main aspect, *Vocabulary and word usage: understanding words and phrases*, is further subdivided into two sub-aspects: *sentence comprehension* and *word comprehension*. Both of these sub-aspects concern the extent to which students are capable of deriving phrase and word concepts from the context. First, it concerns the ability to use compensating strategies to derive meaning at the sentence level. It also has to do with the ability to indicate or derive the meaning of frequent or less frequent words based on word types, word forms, or spelling within a familiar context.

Table 5-6. Student model for English reading comprehension (College voor Toetsen en Examens, 2014)

Main aspect	Sub-aspect
1. Targeted reading: Applying reading strategies (to be tested through the sub-aspects of 2 and 3.)	1.1 Orientation, skimming; scanning, deriving the main idea (3.1, 3.2) 1.2 Selective reading (searching); recognizing relevant details from the text (3.4) 1.3 Intensive reading (2.2, 3.3, 3.5)
2. Recognizing text functions: recognizing the objective and type of the text	2.1 Recognizing the author's objective *Is capable of recognizing the author's objective: Informing, Persuading, Amusing, Instructing* 2.2 Using text information and drawing conclusions *Is capable of using textual information to draw conclusions beyond the literal meaning of the text (e.g., comparison, classification, illustration)*
3. Understanding text structure: extracting information from texts	3.1 Deriving the main idea of the text *Is capable of identifying and/or expressing the most important idea from the text* 3.2 Distinguishing main aspects from side issues *Is capable of distinguishing important information within the text from less important information* 3.3 Identifying key phrases *Is capable of identifying key phrases (e.g. the sentence containing the most important information in a paragraph)* 3.4 Recognizing relevant details *Is capable of searching for detailed information in a text* 3.5 Understanding text structure *Is capable of understanding the structure of a text (e.g., orderings in time, causes)* 3.6 Understanding references between sentences *Is capable of understanding the meaning of reference words* 3.7 Understanding the meaning of signal words *Is capable of understanding the meaning of signal words used in sentences*
4. Vocabulary and word usage: understanding words and phrases	4.1 Sentence comprehension: understanding sentences in context *Is capable of indicating or deriving the meaning of sentences in the context of a text on a familiar topic* 4.2 Word comprehension: understanding words in context *Is capable of indicating the meaning of words (frequent or infrequent) used in a sentence on a familiar topic, based on the type, form, and spelling of words*

5.2.2 Operationalization of English reading comprehension

For reading comprehension, tasks were also developed by a construction group consisting of secondary teachers, divided into two construction groups (one for vmbo and one for havo/vwo). The tasks for reading comprehension were also approved by the confirmation committee.

The student model for reading comprehension (14 sub-aspects) is much more detailed than the student model for writing skills (8 sub-aspects). On the one hand, this resulted in a more targeted construction. For example, the process of constructing based on signal words is much more detailed than is the case for the comparable sub-aspect of writing skills, which assesses the entire use of structure words. On the other hand, it yields less variation within the sub-aspect. The results also indicate that it is difficult to distinguish between some sub-aspects. For example, consider the identification of the main idea and a key phrase. In short texts, these sub-skills overlap. It is important to make good agreements in this regard: what do we understand a given sub-skill to entail? In addition, a pre-test could provide additional information on the utility of such a detailed character.

The first main aspect is intended to measure a student's reading strategies. This would pose a challenge in the current assessment set-up, as it only allows measuring the outcome of a reading task, and not how the reading task is tackled. The construction started with the developing tasks for the other main aspects. During the process of construction, the strategies necessary for solving each developed task were estimated. This revealed that certain tasks clearly called for one particular reading strategy, while in other tasks strategy usage was strongly dependent upon the student. Based on this knowledge, we selected for Main Aspect 1. Targeted Reading several tasks, which apparently did clearly call for a particular strategy (see Table 5-7).

In the context of the various sub-skills as well, some types of tasks proved more suitable than others did. The majority of the tasks were multiple-choice (see Table 5-8 and Table 5-7). This type of task is best suited for the sub-skills calling for a main idea, drawing conclusions about the information from the text or identifying the author's objective, because an answer beyond the text is expected, but it was not possible to evaluate an open-ended task. Sub-skills like identifying key phrases call for action in the actual text, such that the assignment could involve marking in the actual text (i.e., marking tasks). The understanding of text structure can also be assessed in a more active manner. One option would be to have students drag elements within the structure of a text or to select the proper structure word in the actual text (drop-down assignment).

Table 5-7. Number of confirmed tasks and responses, for each sub-aspect of the Student Model for English Reading Comprehension

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
Sub-aspect	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks**	**Number of tasks and responses**
ENG-1.1	8 (8)	10 (10)	3 (6)	5 (5)	8 (8)	33 (36)
ENG-1.2	4 (13)	3 (8)	2 (7)	3 (11)	2 (5)	14 (44)
ENG-1.3	5 (8)	3 (12)	2 (8)	6 (18)	5 (11)	18 (48)
ENG-2.1	10 (10)	12 (12)	6 (6)	11 (11)	9 (9)	46 (46)
ENG-2.2	7 (7)	14 (14)	14 (17)	9 (12)	6 (9)	50 (59)
ENG-3.1	8 (8)	23 (23)	10 (10)	9 (9)	7 (11)	56 (60)
ENG-3.2	8 (8)	10 (10)	6 (8)	11 (11)	9 (9)	43 (45)
ENG-3.3	6 (9)	9 (12)	5 (5)	9 (14)	7 (10)	32 (46)
ENG-3.4	10 (19)	26 (61)	8 (20)	15 (35)	4 (13)	62 (147)
ENG-3.5	8 (8)	5 (8)	4 (9)	9 (11)	6 (9)	32 (45)
ENG-3.6	5 (8)	11 (11)	10 (20)	13 (14)	11 (12)	42 (56)
ENG-3.7	7 (7)	12 (12)	10 (10)	11 (11)	7 (9)	47 (49)
ENG-4.1	9 (9)	10 (12)	13 (13)	11 (11)	15 (15)	45 (47)
ENG-4.2	10 (13)	8 (11)	7 (9)	16 (24)	9 (14)	50 (71)
Total	105 (135)	156 (216)	100 (148)	138 (197)	105 (144)	570 (799)

Table 5-8. Number of confirmed tasks and responses for English Reading Comprehension, by type of task

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**
Paragraph task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Correction task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Drop-down task	2 (6)	6 (17)	6 (14)	10 (10)	9 (11)	29 (51)
Hotspot task	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)
Short open-ended task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Marking task	24 (27)	23 (26)	26 (33)	25 (31)	25 (29)	109 31)
Matrix task	8 (28)	14 (48)	8 (30)	13 (47)	6 (24)	48 (173)
Multiple-choice task	60 (60)	105 (105)	54 (54)	72 (72)	52 (52)	329 (329)
Multiple-response task	1 (1)	0 (0)	0 (0)	0 (0)	0 (0)	1 (1)
Dragging task	1 (4)	3 (12)	4 (15)	11 (30)	11 (26)	29 (86)
Dragging task with image	0 (0)	1 (4)	0 (0)	0 (0)	0 (0)	1 (4)
Ordering task	8 (8)	4 (4)	2 (2)	7 (7)	2 (2)	23 (23)
Total	105 (135)	156 (216)	100 (148)	138 (197)	105 (144)	570 (799)

5.3 References

College voor Toetsen en Examens (2014), *Publieksversie Toetswijzer Diagnostische Tussentijdse Toets voor Nederlands, Engels en Wiskunde [Public Version of the Assessment Specification for the Diagnostic Educational Test for Dutch, English and Mathematics].* Utrecht: College voor Toetsen en Examens. Retrieved from https://www.pilotdtt.nl/documenten/publicaties/2014/12/15/toetswijzer-dtt

Deane, P. (2011). *Writing Assessment and Cognition*. Princeton, New Jersey: Educational Testing Service.

Deane, P., Odendahl, N., Quinlan, Th., Fowles, M., Welsh, C. & Bivens-Tatum, J. (2008). *Cognitive Models of Writing: Writing Proficiency as a Complex Integrated Skill*. Research Report ETS RR-08-55. Princeton, New Jersey: Educational Testing Service.

Meijerink, H. (2008). *Over de drempels met taal en rekenen: hoofdrapport van de Expertgroep doorlopende leerlijnen taal en rekenen [Over the hurdles with language and arithmetic: Main report of the Expert Group on continuing educational streams in language and arithmetic].* Enschede: Expertgroep Doorlopende Leerlijnen Taal en Rekenen.

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

Van Hest, E., Beltman E. H., & Kleintjes, F. (2001). *Koppeling mvt-examens aan het Europese Referentiekader [Linking MVT examinations to the European Frame of Reference]. Arnhem*: Citogroep, unit Voortgezet Onderwijs.

6 Assessment of open-ended writing tasks

Roelien Linthorst and Karen Keune[8]

6.1 Introduction

The DET Assessment Specification (College voor Toetsen en Examens, 2014), which served as the foundation for the development of the DET, stated that supplementary open-ended writing tasks would be developed for both English and Dutch, in addition to the closed-ended, adaptive, and diagnostic parts of the DET. The following was stated in the Assessment Specification in this regard:

> In the open-ended writing task, the student produces a text based on an assignment. The task is completed on a computer. Automated scoring for open-ended writing tasks in the DET is still in development. For this reason, the open-ended writing task will be optional for the first few years. The teacher can assess the student's level of writing skills according to an assessment format. In time, it should be possible to assess the open-ended writing task automatically (College voor Toetsen en Examens, p. 11).

The Assessment Specification committee's choice to want to offer open-ended writing tasks (optional or compulsory) in addition to the short open-ended and closed-ended tasks is related to the necessity of dividing writing skills into subskills in the DET, so that they can be assessed separately in order to arrive at a proper, refined diagnosis. These subskills are nevertheless difficult to distinguish from each other when writing an actual text. Writing is a complex skill, in which a writer is constantly working on a complicated interplay of various sub-activities, as also demonstrated by the model developed by Hayes (1996), in which the writing process is presented (see Figure 6-1).

In open-ended writing tasks, all these sub-processes play a role at nearly the same time. For this reason, open-ended writing tasks correspond well to what takes place in ordinary teaching practice. In spite of this, disadvantages are associated with using open-ended writing tasks as a assessment instrument. As indicated in the Hayes (1996) model presented in Figure 6-1, the act of writing a text calls upon a variety of cognitive processes. The various aspects of writing are often not directly visible in the assessment of open-ended writing tasks. Different students may receive the same mark on a writing task, even though their texts differ on a wide range of points. The weighting of the various aspects in the mark is often unclear. In addition, there is often little agreement between assessors in the case of open-ended writing tasks (De Glopper & Willemsen, 2014). This can be explained in part by differences in the ways in which teachers value various aspects of writing skills. One teacher might assign much greater weight to spelling and formulation errors in the final evaluation than another teacher does. Another disadvantage of open-ended writing tasks is that their administration requires a great deal of time, such that students are often presented with only one writing task in a given test. As demonstrated by Bouwer and Van den Bergh (2015), however, one teacher's assessment of a single writing task does not allow any judgement concerning a student's general writing skills, given the large differences between the assessments of teachers and the large differences between the performance of students on different writing tasks. They argue that a writing test should consist of at least 12 tasks in order to allow generalization across genres.

[8] We are grateful to Rick Godschalk, Hanna Trommel, and Jennifer Leusink for their contributions to the various studies.

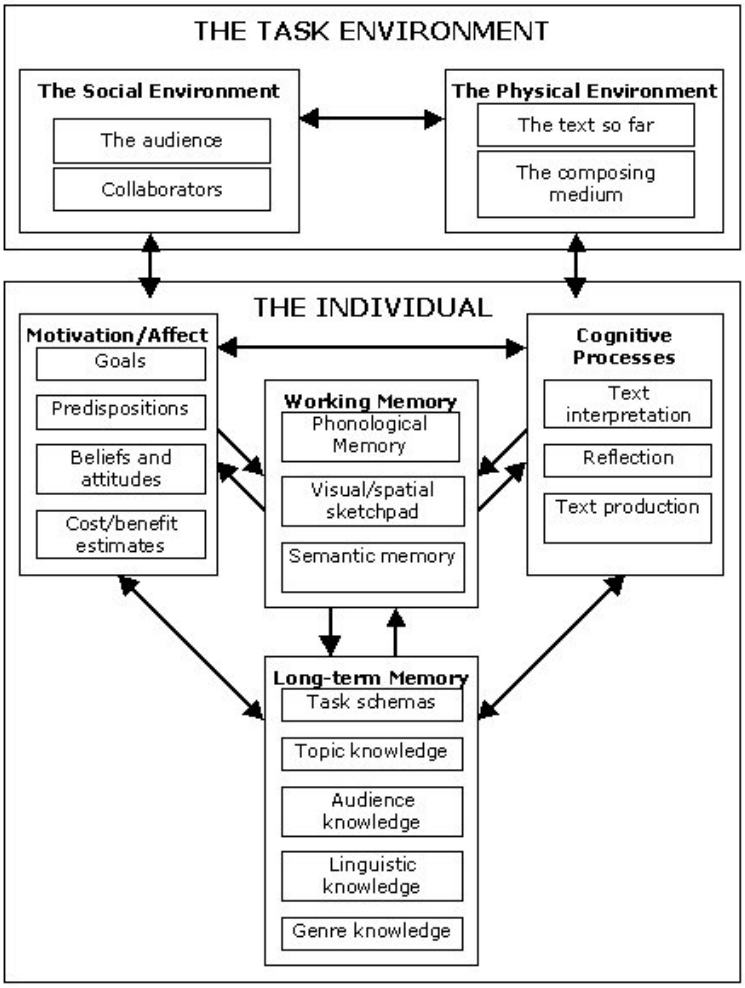


Figure 6-1. Hayes' model of writing skills (1996)

In closed-ended writing tasks, students can be presented with assignments that address only one aspect of writing at a time. This allows the assessment results to provide focused information that students can use in order to improve their writing skills. Such tests also make it possible to present students with a variety of contexts and writing objectives in a short time, thus increasing the efficiency of test administration and doing more justice to the diversity of writing skills than would be possible with a single open-ended writing task. A closed-ended form of assessment for writing skills also lends itself well to adaptive test administration, which is more difficult to realize with open-ended writing tasks (Linthorst & Schuurs, 2014).

Briefly stated, open-ended writing tasks call upon all cognitive processes that are relevant to the act of writing, and they correspond to what takes place in teaching practice. For these reasons, open-ended writing tasks make a valid impression. Open-ended writing tasks are nevertheless associated with various disadvantages, including those due to the use of various assessors in their assessment. The possibility of automated assessment of writing tasks might offer a solution to this assessment problem. Partly for these reasons, the preliminary study of the DET includes a suggestion to opt for a multi-stage model in the final version of the DET. In a multi-stage model, the results on the open-ended writing task/tasks (which will be assessed automatically) would be used in determining the closed-ended adaptive tasks that the student will subsequently complete in order to allow a refined diagnosis (Roelofs & Schouwstra, 2012). Within the framework of the DET, therefore, an exploratory study was conducted into the possibility of assessing student texts automatically. This section consists of a report on this study and the studies that followed it.

6.2 Automated assessment of writing skills

The first pre-test in 2013 (Feenstra & Keune, 2013) included an investigation of the extent to which it would be possible to automate the assessment of open-ended writing tasks. To that date, no applications were available for Dutch. The T-scan legibility predictor is available for Dutch (Kraf, Van der Sloot, Pander Maat, Van den Bosch, Van Gompel, 2013). This instrument is a linguistic-technological application that automatically analyzes texts by dissecting sentences and identifying word types. The application also includes several lists of words and frequencies. Based on this information, T-scan predicts text complexity using more than 150 text attributes.

These text-complexity measures can be used to analyze only lower-order skills. The student model for writing skills in Dutch (see Chapter 4) includes the following aspects: *3 Linguistic skills (word and sentence level)* and *4 Orthographic skills (spelling and punctuation)*. A complete assessment of a writing product is not possible. The objective of the study during the try-out (Feenstra & Keune, 2013) was therefore to determine whether it would be possible to use text attributes to discriminate between writing products of good quality and those of poorer quality according to the aforementioned aspects. If this is the case, it could provide a foundation for the further development of the automated assessment of writing tasks.

For the study, two essay assignments were administered to students from 2 vmbo-bb [basic vocational program, second year]; 2 vmbo-kb [middle-management vocational program , second year]); 2 vmbo-gt [theoretical programs of pre-vocational secondary education , second year]; 3 havo [senior general secondary education , third year]; and 3 vwo [pre-university education , third year]). The 438 writing products by the students from the second year of vmbo were compared to 253 writing products by students from the third year of havo/vwo. For this purpose, text attributes were selected from T-scan that could provide information about two sub-aspects of *3 Linguistic skills (*word usage and sentence structure) and the sub-aspect *2.3 coherence* within the text. The automated spell-checker Valkuil.net was used to measure *Orthographic skills (spelling and punctuation)*.

According to the results of the study, the use of certain measures for word usage, sentence structure and coherence were more characteristic of students in vmbo, while other measures were more characteristic of students in havo/vwo. Closer inspection of individual measures clearly indicated that the reliability of the analyses was strongly influenced by a number of factors. The specific writing task had a stronger influence on the text attributes than did the educational stream of the student. Further, texts with spelling errors, non-existent words and abbreviations were incorrectly assessed a having a high word complexity, and texts with grammatically incorrect sentences and sentences with missing punctuation were incorrectly assessed as having high sentence complexity.

These results indicated that automated assessment of writing products within the DET was not yet possible. Analysis of text attributes provides interesting information about lower-order skills, but not in such a way that a student can be assessed according to this information. For Dutch, further development in this regard is a highly complex, time-consuming, and expensive process, and it therefore does not seem feasible for the foreseeable future. The development of a tool for diagnostic feedback on lower-order skills in writing products does appear feasible, however, and it could be valuable to both students and teachers.

6.3 Research on marking models

As outlined in the previous section, it is not yet possible to conduct complete, reliable assessments automatically. In the pilot, it was thus not possible to offer open-ended writing tasks within the DET that could be marked automatically. It did appear possible to develop open-ended writing tasks accompanied by proven marking models that teachers could use to evaluate the text of a student. Two studies were conducted in this regard. In one study, which was conducted during the second pre-test, three different marking models were examined for their utility for both Dutch (Linthorst & Keune, 2014) and English (Godschalk, Vrijs & Schouwstra, 2014).

In a second study conducted in 2016, new writing tasks were developed based on the experiences gained in the first study, and a new exploration was performed on the association between these open-ended writing tasks and the results on the DET. Both of these studies were conducted for both English and Dutch.

One intended side-effect of the two studies was that they would result in the collection of usable writing products by students, which could be used in the further development of the automated assessment of writing skills. To this end, a corpus of student texts is quite desirable.

6.3.1 Research on marking models for Dutch

The first study, on the utility of the three different marking models, was conducted in the spring of 2014 for both English and Dutch. For Dutch, two open-ended writing assignments were presented to the students (Linthorst & Keune, 2014). The Dutch study had two objectives. First, the student products collected served as potential input for a follow-up study on the automated assessment of writing skills. In addition, the student products were used to assess the reliability and utility of three potentially suitable marking models for inclusion in the DET. The three marking models were as follows:

1. The *global marking model*, in which the marker is asked to render a judgment on the text according to four evaluation points, which correspond to the main aspects of the DET (*objective & audience*, *structure*, *word usage & sentence construction* and *spelling & punctuation*).

2. *Marking using anchor texts*, in which anchor essays are used to mark specific boundaries. In this study, anchors were selected for each of the four main aspects and for vmbo-tl, havo and vwo at the boundary "below & at level", as well as the boundary "at & above level". The markers were then asked to position the texts to be evaluated alongside these anchor essays, as illustrated in Figure 6-2.

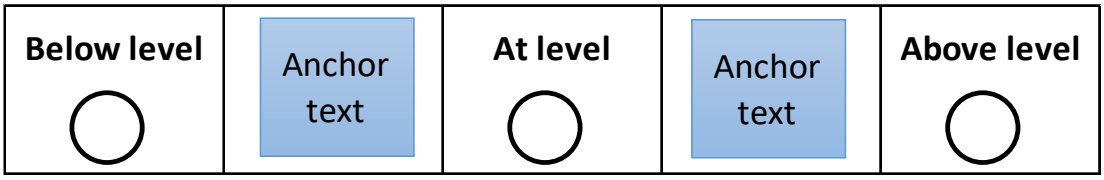


Figure 6-2. Schematic presentation of anchor marking

3. The *analytic marking model*, in which markers were asked to mark a text according to several evaluation points that together produce the most accurate image possible of a student's level of writing skills. The evaluation criteria for this model corresponded to the 12 (vmbo) and 13 (havo/vwo) sub-aspects distinguished in the student model for writing skills.

In all, 480 student texts were collected and marked by 8 markers according to the design illustrated in Table 6-1. For example, Marker 1 marked Texts 1–60 using the anchor model and the analytic model, Texts 181–240 with the global model and the anchor model, and Texts 241–360 with the global model and the analytic model.

Table 6-1. Illustration of marking design for student texts

	Text							
Model	**1-60**	**61-120**	**121-180**	**181-240**	**241-360**	**301-360**	**361-420**	**421-480**
Global	5	3	7	1	1	6	7	4
	4	6	2	8	5	2	3	8
Anchor	1	3	2	4	5	8	7	6
	4	2	3	1	7	6	5	8
Analytic	1	2	7	8	1	2	5	6
	5	6	3	4	7	8	3	4

The marking was done from global to specific. The marker either started by marking the texts according to the global model and then according to the anchor model or the analytic model, or started by marking according to the anchor model and then according to the analytic model. This prevented the marking using the global and anchor models from being influenced by the detailed marking in the analytic model. After

each marking round, the markers were also asked about their experiences with working with the specific marking model. This was done orally, based on a questionnaire that had been prepared in advance. This resulted in the collection of supplementary information on the quality, utility, and efficiency of the models.

A detailed report of the results can be found in Linthorst and Keune (2014). The inter-rater reliability of all models was reasonably good. Inter-rater agreement for the anchor model was somewhat higher on most points than it was for the global model. For this model, the actual markings were also more widely dispersed. No remarkable differences were found between the global model and the analytic model. Although working with the anchor texts appeared to be somewhat better with regard to inter-rater reliability, this method of marking was not preferred by the markers. They found the process of working with anchor texts to be more time-consuming, possibly in part because they were given new anchor texts for each level and each main aspect. In addition, several markers had questions concerning the choice of several anchor texts, which they did not consider appropriate for the level to be assessed. The marking according to the global model proceeded most quickly, and the markers regarded it as the most efficient. According to the markers, the analytic model did the most justice to the performance of the students, as their texts were marked on many different aspects.

6.3.2 Research on marking models for English

Also for English, a study was conducted in 2014 on the suitability of various marking models, although the study had a somewhat different approach (Godschalk, Vrijs and Schouwstra, 2014). The English student products collected were marked according to a marking model that addresses the global quality of the writing product, as well as its communicative effectiveness. Communicative effectiveness concerns whether the student has actually achieved the intended writing goal. Communicative effectiveness is not included in the student model for English, as it is difficult to test with closed-ended tasks. Because the Assessment Specification committee for English considered it important for open-ended writing tasks, however, a study was conducted on the possibility of assessing communicative effectiveness separately in open-ended writing tasks.

In the study, three writing tasks were administered to students in vmbo and havo/vwo. This yielded 571 writing products. The evaluation of both the global quality and the communicative effectiveness of each writing product was conducted holistically. The marker was asked to assign a numerical mark to the writing product, along with a classification of the educational stream corresponding to the writing product. A large share of the student texts (60%) were marked by two markers, with the rest being marked by only one marker.

A detailed report on this study is also included in the 2014 research report on the DET (Godschalk, Vrijs and Schouwstra, 2014). According to the results, the markers were unable to distinguish between communicative effectiveness and the global quality of the writing products.

6.4 Development of open-ended writing tasks and initial exploration of coherence

The experiences gained during the previous study on the automated assessment of writing skills (Section 6.2) and the study on marking models (Section 6.3) gave rise to a follow-up study in the 2016-2017 academic year. This small-scale study had two parts. First, it served to develop several open-ended writing tasks for English and Dutch within the framework of developing the DET. At the same time, it was intended as a small-scale exploration of whether and what type of association existed between the results of students on the DET and on one or more open-ended writing tasks.

6.4.1 Development of open-ended writing tasks

In pursuit of the aforementioned objectives, the English and Dutch construction groups developed several open-ended writing tasks in mid-2016. Important principles in the construction of these tasks for both languages included the following:

- There should be separate writing tasks for vmbo and for havo/vwo. Given the difference in level between the second year of vmbo and the third year of havo/vwo, it was not considered realistic to

administer exactly the same writing tasks to all of these students. Some writing tasks did lend themselves to being administered to students in both vmbo and havo/vwo students, as well as for within-group comparisons. In other cases, however, the various educational streams were differentiated, including at the recommendation of the confirmation committees.

- Students should be able to complete each writing task within a single classroom period of approximately 50 minutes, in order to avoid placing an excessive time burden on the participating schools.

- The writing task should have a clear communicative objective, should be as authentic as possible and should correspond to the life experiences of the student as much as possible.

- Writing tasks should be developed for various text objectives (e.g., informing, persuading) and various text genres (e.g., email, article, instructions) in order to represent the domain of writing as broadly as possible.

- The “copyability” of the writing task should be kept to a minimum. Some instruction for the writing task is desirable, for purposes of comparing the ultimate student writing products. Moreover, students need background information about the topic in order to write a substantively good text. At the same time, however, such instructions increase the “copyability” of the task. Students can often incorporate phrases from the instruction or source material into their own texts, thereby diminishing their production of original text. To address this problem, in the development of the open-ended writing tasks, we sought input that can be provided in another manner (e.g., through audio or video material, or through images). Moreover, for English, the choice was made to provide the assignment in Dutch, so that students would not be able to copy English concepts verbatim from the instruction.

As in the regular DET, the writing tasks that have been developed were submitted to the confirmation committees. They were also submitted to lesson-plan developers from Netherlands Institute of Curriculum Development (SLO) for Dutch and English. Three writing tasks were ultimately identified to use in a small-scale pre-test for both Dutch and English (see Table 6-2).

Table 6-2. Selected writing tasks for Dutch and English

Subject	Writing Task	Educational stream
Dutch	*Gescheiden gymles* (Segregated Gym Class)	vmbo
	Tekenbeet (Tick Bite)	vmbo & havo/vwo
	Voedingsles (Nutrition Class)	havo/vwo
English	Homework	vmbo
	Fantastic Beasts	vmbo & havo/vwo
	Olympic gymnast	havo/vwo

Based on the experiences gained in the first study (see Section 6.3), the choice was made to provide a global marking model along with the writing tasks, in which the markers are expected to mark the tasks according to the four main aspects from the student model for writing skills. Although the results from the first study identified the model with anchor texts as the most promising with regard to reliability and inter-rater agreement, this model also proved labor-intensive. In the second study, we were not able to request much time for the marking of the student texts. We wanted to use experienced teachers for the marking of the student texts, and a great deal of time was already being asked of the participating schools, as they administered the open-ended writing tasks in addition to the regular DET. For this reason, and due to the fact that the use of anchor marking would require an additional administration in order to collect potentially suitable anchor essays, the choice was made to use a global marking model. The marking model for Main Aspect 1 for Dutch is presented in Figure 6-3.

1. Coordination to objectives and audience (rhetorical skills)
Is the student capable of writing a targeted text tuned to the requested audience?

Aspects that play a role in this regard include the following:
Can the student estimate which information should and should not be included?
Can the student select and maintain the proper tone?
Can the student adopt and maintain a suitable writing objective?
…compared to what a student in this educational stream (vmbo-bb, vmbo-kb, vmbo-gt/TL, havo or vwo) should be capable of doing

Very weak	Weak	Moderate	Satisfactory	Good	Very good		Not markable
1	2	3	4	5	6		0

Figure 6-3. Global assessment model for Main Aspect 1 for Dutch

As shown in the figure, the marking model includes a six-point scale ranging from very weak to very good. The marker is expected to indicate how well the student has performed on the main aspect in comparison to what a student in the same educational stream should be capable of doing. The presentation of the sub-aspects along with the main aspect is intended to make clear how the main aspect is defined within the DET.

6.4.2 Exploration of coherence in open-ended writing tasks and DET

In addition to the development of several open-ended writing tasks, we wanted to conduct an initial exploration of the association between texts written by students and their results on the adaptive DET. We also wanted to see how the writing tasks were carried out in practice and gain experience with the administration of open-ended writing tasks in Facet. One problem in the administration of the open-ended writing tasks was that multiple marking (the marking of a student product on multiple aspects) was not yet possible in Facet, and we wanted to have the writing products marked on the four different main aspects. To address this problem, we ultimately chose a work-around solution, in which the student was required to click through an empty screen three times after writing the text, so that the marker would ultimately be able to mark the student product on four separate screens (one for each of the four aspects).

In order to conduct a pilot exploration of the open-ended writing tasks and their association with the closed-ended DET, the schools administering the DET were approached in the spring of 2017 to ask if they would be willing to administer an open-ended writing task for English and/or Dutch in addition to the regular DET in one or more classes. One requirement for participation was that the school would also be expected to evaluate the writing products of its students based on the marking model included with the writing tasks. In addition, the writing products would be marked by a second, independent marker – a Cito intern from Radboud University Nijmegen (for English) and Cito interns from the University of Groningen (for Dutch). Given that the marking of open-ended writing tasks often differs depending on who marks, the marking of each writing product by two individuals was a minimum condition for this study.

The school's own marking proved to be a major obstacle to participation. The response for participation in this study was low. The substantial time investment that schools had already made in the DET also played a role, such that many schools perceived the administration of an additional assignment as an excessive burden. The number of schools registering to administer the open-ended writing tasks is presented in Table 6-3, along with the number of administrations that actually took place.

Table 6-3. Participation in the open-ended Writing Task in Spring 2017

	Number of vmbo **schools registered**	**Number of havo/vwo schools registered**	**Actual participation**
Dutch	2	1	2 vmbo schools with a total of 175 students and 1 havo/vwo school with a total of 31 students
English	1	1	1 havo/vwo school with a total of 27 students

Three schools ultimately administered an open-ended writing task for Dutch, with one school participating in the writing task for English. Unfortunately, the schools participating for Dutch ultimately decided not to

conduct their own marking of the open-ended writing task. We are therefore not able to report on the results for Dutch. In the following section, therefore, we report only on the results for English.

The open-ended writing task for English was administered to 27 students in the third year of havo. The “Fantastic Beasts” assignment consisted of writing a YouTube post based on a video. In this assignment, students were asked to indicate whether they would like to see the film described in the video, stating at least three arguments. These instructions were given in Dutch, in order to prevent verbatim copying. The video was presented in English.

The 27 student products were marked by two markers: the students’ English teacher and a Cito intern from Radboud University Nijmegen. Based on their marking, we started by examining the extent of agreement between the markers. Unfortunately, there was too little agreement between the markers. The low agreement between markers indicates that the markings are unstable and unsuitable for use in further analysis. These data unfortunately do not allow any exploration of the association between the open-ended writing task and the results on the actual DET.

One likely reason behind the disappointing level of agreement between markers for the open-ended writing task in English is that one of the markers was not a teacher. In the earlier study on the various marking models, we observed a relatively good level of inter-rater agreement. In that study, however, only teachers with classroom experience and who had also attended a group training session were involved. That was not the case in this study.

6.5 Conclusions and recommendations

Within the three-year DET pilot study, it was not possible to offer writing tasks that could be reliably assessed automatically. The two studies that we conducted yielded useful insight into the administration and marking of open-ended writing tasks.

With regard to the *marking* of open-ended writing tasks by markers, the first study indicated that anchor marking is promising, albeit time-consuming, particularly if texts from specific educational streams must be marked on multiple aspects (with differing anchors). Setting up the anchor marking also requires a substantial time investment, as it should ideally include an initial administration of the writing task in order to select the anchors before they can actually be used. Global marking based on several aspects is therefore much more practical, although the level of inter-rater agreement in the second study was disappointing. We were thus unable to identify the ideal marking model in these two studies. Recent developments involving pair-wise comparisons (e.g., Van Daal, Lesterhuis, Coertjens, Donche, & De Maeyer, 2016) might offer possibilities for arriving at a reliable marking in a fairly efficient manner. Pair-wise comparison requires a considerable investment of time, however, and it ultimately yields a ranking of student texts, thus possibly offering fewer links for a formative type of writing-skills evaluation.

With regard to the actual *writing task*, the second study demonstrated that providing source material through audio/visual fragments offers clear advantages relative to a written case. It reduces copyability, and the school visits that we conducted revealed that students find audio or visual material more appealing. This offers prospects for the future with regard to the further integration of language skills in language-skills assessment.

It will also be necessary to examine how the instructions for the writing tasks can be given to the students. The schools participating in the administration of the writing tasks received a manual outlining the administration procedure and indicating which instructions were to be given to the students. In practice, however, the teachers did not follow these instructions to any great extent. The instructions were largely limited to the instruction that students could log in with their login information and start working on the assignment. Reinforcing the instructions in another manner, e.g. through an audio or video post in the administration environment, therefore seems advisable. This would probably ensure that the students would receive better, more consistent instructions.

In any case, it appears advisable for open-ended writing tasks to be accompanied by training for markers. In these studies, the experience and training of markers appeared to have had an effect, as the agreement

between the markers in the first study was clearly higher than it was in the second study. In the first study, a group starting session was organized for the markers, in which the global marking principles were discussed. In the second study, the markers did not receive any specific instructions in advance.

The desired comparison between open-ended writing tasks and the results on the DET proved impossible in the second study. Proper implementation will require a more extensive research, preferably with the administration of more varied open-ended writing tasks to the same students, given the task-dependence of writing skills (e.g. Brouwer & Van den Bergh, 2015; Feenstra & Keune, 2013). A more qualitative comparison also seems advisable, as the results on an open-ended writing task are not immediately comparable to the results on the DET. In light of the discussions concerning the open-ended and closed-ended assessment of writing skills (see e.g., De Glopper & Willemsen, 2014; Pulles, Den Ouden, Herrlitz, & Van den Bergh, 2013), further research on open-ended and closed-ended forms of writing-skills assessment certainly merit recommendation for the future.

6.6 References

Bouwer, R. & Van den Bergh, H. (2015). Toetsen van schrijfvaardigheid: hoeveel beoordelaars, hoeveel taken? [Testing writing skills: How many assessors, how many tasks?] *Levende Talen Tijdschrift, 16* (3), 3-12.

College voor Toetsen en Examens (2014), *Publieksversie Toetswijzer Diagnostische Tussentijdse Toets voor Nederlands, Engels en Wiskunde [Public Version of the Assessment Specification for the Diagnostic Educational Test for Dutch, English and Mathematics].* Utrecht: College voor Toetsen en Examens. Retrieved from https://www.pilotdtt.nl/documenten/publicaties/2014/12/15/toetswijzer-dtt

De Glopper, K., & Willemsen, A. (2014). Kunnen gesloten toetsen bijdragen aan de toetsing van schrijfvaardigheid [Can closed-ended tests contribute to the testing of writing skills]? *Levende Talen Tijdschrift*, 15(1), 31-38.

Feenstra, H. & Keune, K. (2013). Geautomatiseerde beoordeling schrijfvaardigheid Nederlands [Automated assessment of Dutch writing skills]. In S. Schouwstra (Ed.), *Diagnostische tussentijdse toets: Onderzoek 2013 [Diagnostic Educational Test: Research 2013]* (pp. 63 - 91). Arnhem: Cito.

Godschalk, R., Vrijs, W. & Schouwstra, S. (2014). Beoordeling open vragen schrijfvaardigheid Engels [Assessment of open-ended tasks on English writing skills]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [The Diagnostic Educational Test: Research 2014]* (pp. 82-90). Arnhem: Cito.

Hayes, J.R. (1996). A New Framework for Understanding Cognition and Affect in Writing. In: C.M. Levy, S. Randell (Ed.). *The Science of Writing. Theories, Methods, Individual Differences and Applications.* Mayhwah, New Jersey: Lawrence Erlbaum.

Kraf, R., van der Sloot, K., Pander Maat, H., van den Bosch, A., & van Gompel, M. (2013). *T-Scan*, http://languagelink.let.uu.nl/tscan.

Linthorst, R. & Keune, K. (2014). Beoordeling open vragen schrijfvaardigheid Nederlands [Assessment of open-ended tasks on Dutch writing skills]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [The Diagnostic Educational Test: Research 2014]* (pp. 71-82). Arnhem: Cito.

Linthorst, T.R. & Schuurs, U. (2014). Diagnostische toetsing bij het schoolvak Nederlands [Diagnostic assessment for the school subject Dutch]. *28ste Conferentie Onderwijs Nederlands, 28*, 59-63.

Pulles, T., Den Ouden, H., Herrlitz, W., & Van den Bergh, H. (2013). Kan een meerkeuzetoets bijdragen aan het meten van schriftelijke taalvaardigheid? [Can a multiple-choice test contribute to the measurement of written language skills?] *Levende Talen Tijdschrift*, *14*(2) 31-41.

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

Van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V. & De Maeyer, S. (2016). Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in education: principles, policy & practice.* http://dx.doi.org/10.1080/0969594X.2016.1253542.

7 DET Mathematics

Irene van Stiphout

7.1 Student model

The value of the DET lies in the fact that it offers an explanation for why the performance of some students lags behind. The idea for this emerged from the findings of various studies (Bruin-Muurling, 2010; Kraemer, 2011; Van Stiphout, 2011; Roorda, 2012) that student levels lag behind expectations. These studies generated the image that a student's mastery is vulnerable: although students might have a reasonable level of mastery with standard tasks, they are not very flexible outside this context. According to these authors, students also tend not to recognize the same mathematical concepts in different subjects, and they tend to make little or no use of knowledge about mathematical concepts that they have acquired in one subject in the context of other subjects (Roorda, 2012). Furthermore, students often use strategies that are not very sophisticated, and their tendency to adhere to these strategies impedes them from further generalization or abstraction (Kraemer, 2011). It was hoped that a framework corresponding to recent developments in teaching could help in this regard.

The action plan Action Plan for Better Performance (*Actieplan Beter Presteren,* Van Bijsterveldt-Vliegenthart, 2011) is aimed at helping students achieve higher levels. This combination of the need for a higher level and insight into why this higher level is not being achieved generated a two-dimensional explanatory student model consisting of a subject-domain dimension and a "subject-specific didactic" dimension (see Table 7-1). The rationale behind the second dimension is that teachers know from their own assessment which topics (i.e., domains) students have or have not mastered, but that they have less insight into possible reasons *why* the intended levels have not been achieved.

In the subject-specific didactic dimension of the student model, a choice was made to adopt a three-way categorization in which the following developments are intertwined:

- Recognizing underlying mathematical structures (Devlin, 2012);
- Attaining a higher level in terms of an (mathematical) object perspective and the ability to cope with multiple meanings (e.g., Sfard, 1991);
- Seeing associations between mathematical concepts within and between the domains of the interim objectives (NRC & MLSC, 2001).

This three-way categorization was first described by Bruin-Muurling (2010), and Van Stiphout (2011) was the first to apply it to the analysis of student results. It has not previously been used as a aspect of a student model. For a more detailed description of the student model, see the preliminary study (Roelofs & Schouwstra, 2012) and the Assessment Specification (College voor Toetsen en Examens, 2014).

Table 7-1. Student model for mathematics (College voor Toetsen en Examens, 2014)

	Subject-specific didactic aspects		
Subject domains	Structure	Ambiguity	Coherence
Domain B: Numbers (and variables)	B1	B2	B3
Domain C: Relationships	C1	C2	C3
Domain D: Measurement and geometry	D1	D2	D3
Domain E: Associations and formulas	E1	E2	E3
Domain F: Information processing and uncertainty (only for havo/vwo)	F1	F2	F3

The innovative character of the student model became evident early in the process of designing the DET. The expectation that diagnoses based on this three-way categorization would indeed be explanatory was quite ambitious. For this reason, in 2012, an exploratory study was conducted among three classes in a school. The central question of this study concerned the extent to which the elaboration of a student could be related to the underlying aspects (recognizing structure, object formation, seeing associations) that tasks were intended to measure. A small set of tasks was presented to a limited number of students. Based on the written answers of these students, supplementary interviews were conducted with students with regard to their strategies. The mistakes that students made seemed indeed traceable to subject-specific didactic aspects.

Several problems with the student model emerged in the course of 2013. For example, test experts were unable to arrive at a satisfactory level of inter-rater reliability in categorizing the tasks into *structure*, *ambiguity*, or *association*. At the same time, some of the tasks that had been developed by the members of the construction groups were “too flat” to fit into the categorization. Furthermore, concerns existed relating to the feasibility of achieving the benefits of the three-way categorization. Finally, the subject-specific three-way didactic categorization could not be identified in the statistical analyses of student responses.

The mathematics assessment experts divided the existing tasks according to this three-way categorization multiple times, in order to arrive at a description of the categorization in the three subject-specific didactic aspects and to enhance the inter-rater reliability of the categorization. In late 2015 and early 2016, considerations were made with regard to the possibility of combining the student model with psychometric perspectives, with the goal of arriving at a student model supported by empirical evidence. Unfortunately, this combination failed to provide any empirical support for the existing student model, nor did it reveal any suggestions for an alternative student model.

Various alternatives were investigated, with the goal of arriving at a model that would combine the positive aspects of the frameworks, that would respond to their limitations and points of criticism, and that would be practicable throughout the entire process – from the construction groups to the final reporting. It was also hoped that it would be relatively easy to link the alternative student model to categorizations applied by schools, including RTTI (reproduction, training – application Level 1, transfer – application level 2, understanding) and OBIT (recall – comprehend – integrate – apply). This was expected to make the reports more meaningful, as teachers would be able to link them to their own teaching practice. These efforts resulted in the following alternative three-way categorization: knowledge – insight – application (KIA). This categorization is comparable to the TIMSS (Trends in International Mathematics and Science Study) student model, in which the three categories are knowing, applying, and reasoning (Mullis & Martin, 2014). Despite several apparent advantages of this model, there was no time or space to elaborate the model further within the three-year DET pilot project.

Number of tasks

The number of tasks that were constructed and approved following administration are presented in Table 7-2. On average, there are 12 tasks that together yield 18 responses for each sub-aspect in a given educational stream. Slightly more tasks were included for each sub-aspect for vmbo (on average, 14 tasks yielding 20 responses) than was the case for havo/vwo (on average, 10 tasks and 15 responses). Of all tasks, 73% (70% of the responses) were directed toward only one educational stream, with the other 27% of the tasks being suitable for two or three educational streams.

Table 7-2. Number of rejected tasks and responses following the 2017 adaptive administration, by sub-aspect of the student model

	vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
B1	26 (28)	22 (24)	15 (24)	10 (21)	11 (17)	75 (97)
B2	15 (32)	15 (34)	14 (19)	10 (15)	11 (19)	59 (106)
B3	14 (18)	12 (16)	13 (15)	8 (22)	9 (24)	40 (61)
C1	20 (25)	18 (23)	18 (22)	10 (10)	9 (10)	63 (71)
C2	12 (20)	12 (20)	18 (27)	9 (11)	8 (10)	32 (43)
C3	11 (18)	13 (17)	16 (16)	12 (20)	7 (15)	45 (61)
D1	17 (23)	21 (22)	15 (16)	10 (14)	10 (14)	54 (66)
D2	15 (32)	13 (24)	14 (20)	8 (14)	8 (14)	45 (85)
D3	10 (14)	12 (16)	11 (14)	10 (12)	13 (15)	34 (45)
E1	12 (14)	16 (24)	15 (21)	12 (16)	14 (17)	55 (76)
E2	8 (8)	12 (13)	11 (12)	12 (20)	9 (10)	33 (43)
E3	11 (17)	9 (15)	10 (19)	5 (11)	5 (11)	34 (61)
F1				12 (16)	9 (13)	14 (18)
F2				10 (16)	11 (18)	15 (22)
F3				9 (13)	7 (11)	11 (15)
Total	171 (249)	175 (248)	170 (225)	147 (231)	141 (218)	609 (870)

7.2 Operationalization

The total numbers of approved tasks are presented in Table 7-3, by type of task. The tasks were constructed by four construction groups, two for vmbo and two for havo/vwo. Each of the construction groups consisted of three teachers who were currently working in the type of secondary education for which they were to construct tasks. Descriptions of each sub-aspect were constructed according to the two dimensions – the substantive dimension (subject domains) and the subject-specific didactic dimension. The members of the construction groups used these descriptions to construct tasks. Nearly all of the tasks were submitted to the confirmation committee (see Chapter 2). Following the pre-tests, the tasks that stood out in the analyses were again submitted to the confirmation committee. This resulted in the tasks being retained, adjusted, or rejected.

It took longer for some tasks to be submitted to the confirmation committee, due to technical problems in the tasks, particularly in the early stages of the process. The wishes of the construction groups regarding the digital tasks were not always in line with the technical possibilities. Digital possibilities increased considerably throughout the construction process, in part due to the emergence of the Digital Mathematics Environment (DME), Geogebra, and the Maxima computer algebra system, which was used for automated assessment.

Several available types of tasks (see Chapter 2) were developed especially for the languages, and they were not used for the development of tasks for mathematics: paragraph tasks, correction tasks, and marking tasks. Multiple-response tasks were used less frequently for the mathematics assessment. Analysis of the pre-test revealed this type of task to be problematic, as they did not necessarily require students to choose an answer. Although not making a choice is also an answer, it does not necessarily provide a proper reflection of a student's level of skill.

Multiple-choice tasks and short open-ended tasks were by far the most commonly used types of tasks. At the start of the construction process, these types of tasks had been the most suitable for the operationalization of the student model and the technological possibilities that were available at that time. The tasks of the Digital Mathematics Environment and the use of the computer algebra system became available starting in mid-2015. In the course of 2016, it became possible to construct Geogebra tasks and to include them in the assessment. These developments are described in greater detail in Chapter 8.

Table 7-3. Number of approved tasks and responses following the 2017 adaptive administration, by type of task

	vmbo-bb	vmbo-gt	vmbo-kb	havo	vwo	Total unique
	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses	Number of tasks and responses
Categorization task	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	1 (1)
Combination task	5 (8)	5 (10)	6 (13)	3 (5)	5 (9)	16 (30)
Drop-down task	16 (26)	9 (14)	12 (18)	12 (33)	8 (26)	44 (89)
DME task	1 (1)	1 (1)	1 (1)	1 (1)	1 (1)	2 (2)
Geogebra task	0 (0)	3 (3)	1 (1)	2 (2)	2 (2)	6 (6)
Hotspot task	2 (2)	2 (2)	2 (2)	0 (0)	1 (1)	7 (7)
Short open-ended task	77 (97)	74 (88)	79 (105)	55 (64)	54 (61)	261 (316)
Matrix task	12 (46)	13 (42)	10 (33)	18 (38)	15 (38)	50 (140)
Multiple-choice task	42 (42)	54 (54)	47 (48)	37 (37)	37 (37)	162 (163)
Multiple-response task	2 (2)	1 (1)	3 (3)	5 (11)	6 (12)	13 (24)
Dragging task	6 (8)	3 (5)	5 (9)	9 (28)	5 (16)	23 (50)
Dragging task with image	5 (14)	1 (1)	3 (9)	3 (10)	4 (12)	11 (29)
Ordering task	3 (3)	3 (3)	6 (6)	2 (2)	3 (3)	13 (13)
Total	171 (249)	170 (225)	175 (248)	147 (231)	141 (218)	609 (870)

The numbers of rejected tasks and responses for each sub-aspect are listed in Table 7-4. In general, it can be concluded that the tasks that were constructed were somewhat difficult for the levels for which they were intended. Attempts were made to adjust this during the course of the construction process. The necessity of constructing easier tasks combined with the exploration of expanding technological possibilities placed considerable pressure on the construction groups.

The numbers of rejected tasks and responses for each type of task are listed in Table 7-5. It is interesting to note that none of the Geogebra tasks were rejected, despite the fact that not all students are accustomed to this type of tasks (see also Chapter 8).

Table 7-4. Number of rejected tasks and responses following the 2017 adaptive administration, by sub-aspect of the student model for mathematics

	vmbo-bb	vmbo-gt	havo	vmbo-kb	vwo	Total unique
Sub-aspect	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**
B1	4 (4)	6 (6)	3 (7)	4 (4)	2 (6)	18 (22)
B2	5 (5)	9 (11)	5 (13)	6 (9)	4 (9)	25 (38)
B3	2 (6)	6 (13)	6 (17)	4 (11)	4 (15)	14 (32)
C1	1 (1)	3 (3)	5 (9)	6 (6)	3 (7)	14 (18)
C2	3 (3)	4 (4)	4 (5)	4 (4)	3 (4)	8 (9)
C3	6 (6)	4 (5)	3 (4)	8 (8)	4 (5)	16 (18)
D1	6 (6)	6 (6)	5 (5)	6 (6)	4 (4)	18 (18)
D2	2 (2)	9 (14)	3 (3)	2 (5)	2 (2)	15 (23)
D3	5 (5)	6 (6)	5 (6)	5 (5)	3 (4)	14 (15)
E1	2 (2)	5 (10)	4 (9)	4 (5)	1 (1)	14 (24)
E2	4 (4)	4 (4)	2 (2)	4 (4)	4 (5)	10 (11)
E3	4 (8)	5 (12)	9 (17)	5 (10)	8 (11)	24 (43)
F1			2 (4)		2 (4)	2 (4)
F2			3 (3)		0 (0)	3 (3)
F3			5 (8)		4 (7)	5 (8)
Total	44 (52)	67 (94)	64 (112)	58 (77)	48 (84)	200 (286)

Table 7-5. Number of rejected tasks and responses following the 2017 adaptive administration, by type of task for mathematics

	vmbo-bb	vmbo-gt	havo	vmbo-kb	vwo	Total unique
	Number of tasks and responses	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**	**Number of tasks and responses**
Categorization task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Combination task	3 (7)	2 (3)	2 (2)	3 (6)	1 (1)	8 (13)
Drop-down task	1 (1)	3 (4)	3 (14)	2 (5)	2 (10)	8 (23)
DME task	0 (0)	1 (1)	0 (0)	1 (1)	0 (0)	1 (1)
Geogebra task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Hotspot task	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
Short open-ended task	17 (17)	26 (30)	34 (52)	30 (34)	26 (41)	100 (123)
Matrix task	1 (5)	7 (28)	10 (25)	3 (12)	5 (17)	18 (57)
Multiple-choice task	18 (18)	23 (23)	10 (10)	16 (16)	8 (8)	50 (50)
Multiple-response task	1 (1)	1 (1)	3 (6)	1 (1)	4 (4)	6 (9)
Dragging task	0 (0)	1 (1)	0 (0)	0 (0)	0 (0)	1 (1)
Dragging task with image	1 (1)	1 (1)	2 (3)	1 (1)	2 (3)	3 (4)
Ordering task	2 (2)	2 (2)	0 (0)	1 (1)	0 (0)	5 (5)
Total	44 (52)	67 (94)	64 (112)	58 (77)	48 (84)	200 (286)

7.3 References

Bruin-Muurling, G. G. (2010). *The development of proficiency in the fraction domain: Affordances and constraints in the curriculum* (PhD Thesis). Eindhoven: Eindhoven University of Technology. Also available at: https://pure.tue.nl/ws/files/2840528/692951.pdf

College voor Toetsen en Examens (2014), *Publieksversie Toetswijzer Diagnostische Tussentijdse Toets voor Nederlands, Engels en Wiskunde [Public Version of the Assessment Specification for the Diagnostic Educational Test for Dutch, English and Mathematics].* Utrecht: College voor Toetsen en Examens. Retrieved from https://www.pilotdtt.nl/documenten/publicaties/2014/12/15/toetswijzer-dtt

Devlin, K. (2012). *Introduction to Mathematical Thinking*. Petaluma, CA: Devlin.

Kraemer, J-M. (2011). *Oplossingsmethoden voor aftrekken tot 100 [Solution methods for deducting to 100]* (PhD Thesis). Arnhem: Cito. An English summary can be found in the digital version of the thesis at www.alexandria.tue.nl/extra2/721544.pdf

Mullis, I. V., & Martin, M. O. (2014). *TIMSS Advanced 2015 Assessment Frameworks*. International Association for the Evaluation of Educational Achievement. TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement.

National Research Council, & Mathematics Learning Study Committee. (2001). *Adding it up: Helping children learn mathematics*. National Academies Press.

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

Roorda, G. (2012). *Ontwikkeling in verandering: ontwikkeling van wiskundige bekwaamheid van leerlingen met betrekking tot het concept afgeleide [Development in change: Developing the mathematical competence of students with regard to the concept of derivation].* University of Groningen Library.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*(1), 1–36.

Van Bijsterveldt-Vliegenthart, M. (2011, May 5th). *Actieplan beter presteren [Action Plan for Better Performance].* Retrieved from https://www.rijksoverheid.nl/documenten/kamerstukken/2011/05/23/actieplan-vo-beter-presteren.

Van Stiphout, I. M. (2011). *The development of algebraic proficiency* (PhD Thesis). Eindhoven: Eindhoven University of Technology. Also available at: http://alexandria.tue.nl/extra2/719774.pdf

8 Automated assessment of mathematics

Paul Drijvers and Joke Hofstee

8.1 Problem statement

The DET is an adaptive test: it adjusts itself to the level of the student in order to chart the student's strengths and weaknesses in an efficient manner. To make this possible, it is necessary for student responses to be assessed automatically. Automated assessment is thus a condition for adaptivity.

Although this is receiving considerable international interest (see e.g., Fife, 2011; Stacey & Wiliam, 2013; Williamson, Mislevy & Bejar, 2006) and many developments are taking place in this area, we must conclude that no ready-made solutions were available for the automated assessment of mathematics tests. The realization of proper automated assessment was therefore one of the greatest challenges for the DET Mathematics. This challenge was concentrated on three key points: (1) the automated assessment of tasks that allow students space for construction, (2) the "intelligence" of automated assessment, and (3) the assessment of intermediate steps. Additional details on these three key points are provided below.

1) *Automated assessment of tasks that allow students construction space*
   In traditional written mathematics tests, students are able to combine and integrate simple text, formulas, charts, and calculations in their answers, in addition to making sketches or trial calculations on scratch paper. Such space for "doing the math" is usually limited in digital testing environments, even though it is highly relevant for diagnostic assessment. The challenge thus also involves designing tasks for digital mathematics tests in which students can create and experience a construction space in the same way as in paper tests. This requires the availability of mathematical tools, and it imposes specific demands on automated assessment.

2) *The "intelligence" of automated assessment*
   As is the case in other subjects, the proper assessment of responses in mathematics is a subtle matter. For some tasks, the answers x^2-9 and (x-3)(x+3) should both be considered correct, while only the first or the second should be considered correct for tasks that call for elaboration or factorizing. Such decisions are likely to be clear to a human assessor. In automated assessment, it is important to implement mechanisms for such subtleties, along with options allowing the task constructor to direct this for each task.

3) *The assessment of intermediate steps*
   In mathematics, many tasks call for step-wise solutions. In each step of this process, students may make mistakes that will then carry through to the other steps. In such a case, a human assessor (e.g., a mathematics teacher) would usually assign a partial score for the work. Such assessments can be important, particularly for diagnostic purposes. In the automated assessment of digital responses, however, the assessment of intermediate steps is still in the early stages, although there is a demand for it among teachers in the field.

8.2 Working methods

Efforts to realize the automated assessment of mathematics tasks were carried out in close collaboration between assessment experts from Cito, software developers and architects from Cito and from chain partners (e.g., the Board of Tests and Examinations [CvTE] and the Executive Agency for the Department of Education [DUO], which are responsible for the development of the Facet administration environment). The following milestones can be identified in this continuous process of development:

2014: The internal Cito study "Digital diagnostic assessment of mathematics: A study of a new environment" [Digitale diagnostische toetsing van wiskunde: onderzoek van een nieuwe omgeving]. This study was conducted within the framework of the internal research agenda of the Cito department of central exams. In this study, a variant of the DET was developed within Utrecht University's Digital

Mathematics Environment, which offered more options than the QuestifyBuilder - Facet chain did at that moment. This variant was administered to 578 students in the third year of havo/vwo. The conclusion is that environments like Digital Mathematics Environment make it possible to design rich, varied tasks, due to the availability of such mathematical tools as a formula editor, a graphics screen, and a geometry screen. These tools provide students with considerable construction space, thereby allowing more nuanced diagnoses (see Drijvers, Visser, Straat & Schouwstra, 2014).

2015: International expert meeting on automated assessment.
This two-day conference was organized in Arnhem by Cito, in cooperation with the Board of Tests and Examinations. Six leading national and international experts presented the current state of affairs with regard to the automated assessment of digital arithmetic and mathematics tests to 50 participants. Emphasis was placed on the assessment of intermediate steps. The conclusions were that there is a need for built-in tools for charts and geometry in Facet, and that the need for the assessment of intermediate steps is greater for diagnostic tests (e.g., the DET) than it is for summative tests (e.g., the Rekentoets VO [Secondary School Arithmetic Test]) (Drijvers & van Reeuwijk, 2015).

2015-2016: Extending QuestifyBuilder and Facet.
The most important extensions of QuestifyBuilder and Facet were carried out in the period 2015-2016, and they concerned the development of a formula editor (native to Facet), the integration of a module of the Digital Mathematics Environment for drawing charts, the integration of the GeoGebra geometry package, and the use of the Maxima computer algebra system for the automated assessment of formulas and algebraic expressions. These extensions have also led to adjustments of the options that allow task constructors to use these new tools within the QuestifyBuilder author environment.

2016: Interaction study.
This study was an investigation of the interactions of students with the new tools in Facet. The interactions of students with Facet were analyzed based on a detailed observation of 19 students from the third year of havo, vwo and/or vmbo-gt who work with the DET, with particular attention to the new possibilities of the testing environment. This study generated design guidelines for tasks using the Digital Mathematics Environment or GeoGebra (Groenheiden, 2016).

2016-2017: Participation in the Erasmus+ project Advise-Me.
In a consortium with the Open University, Utrecht University, the Université Paris-Est Créteil, and Saarland University, an Erasmus+ research proposal was submitted with the title, “Automatic Diagnostics with Intermediate Steps in Mathematics Education.” This project was awarded, and it was launched in the fall of 2016. It focuses particularly on the automated assessment of intermediate steps. Although this study was not part of the DET pilot project, and although the results will not be available until after the DET project, it emerged in response to a need on the part of the DET.

8.3 Results

The results of the development of automated assessment are discussed according to the three key points specified in 3.1.1. The overview of QuestifyBuilder and Facet, as illustrated in Figure 8-1, can serve as a guideline in this regard.

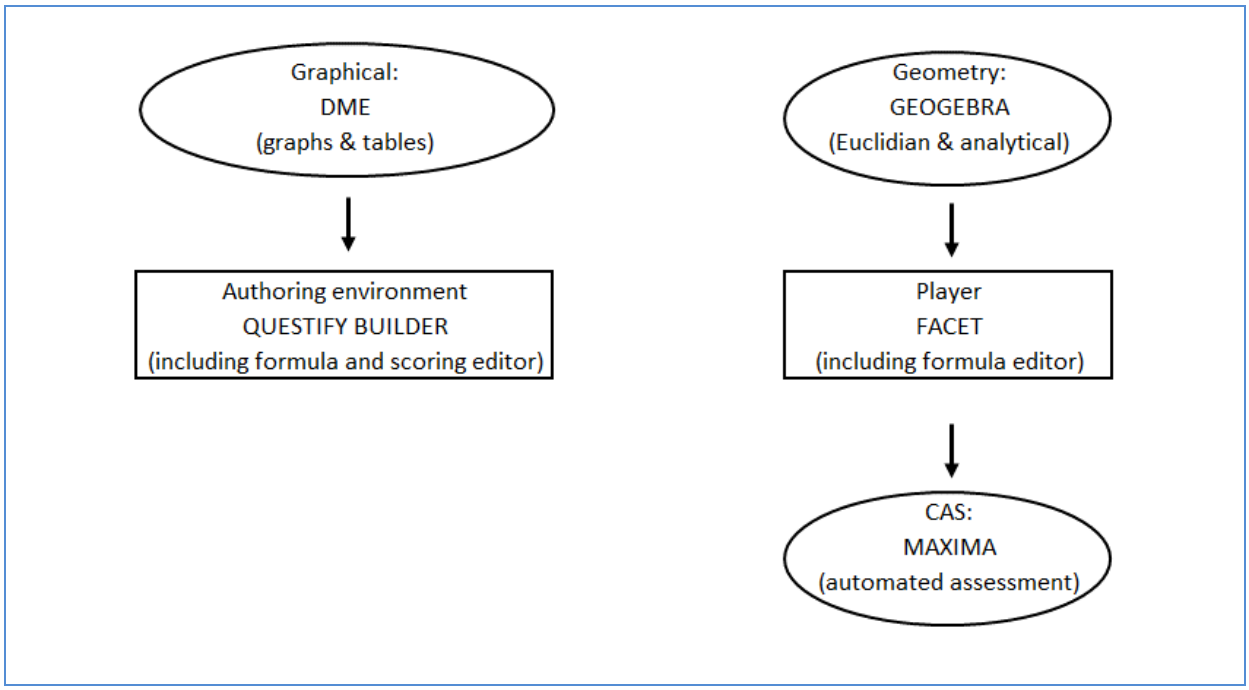


Figure 8-1. Overview of mathematical tools in the architecture of QuestifyBuilder and Facet

The first key point concerns the ***automated assessment of tasks in which students are provided with construction space***. The following was achieved regarding this point.

Entering formulas

A formula editor has been built into Facet, which students can use to enter mathematical formulas and expressions. The formula screen can consist of one or more lines. The formulas entered are interpreted in a standard format (MathML), such that they are also suitable for automated assessment. When designing tasks, task constructors can make the entire formula editor available, although they may also choose limited or minimal variants, which are more suitable for such assessments as the DET in vmbo or the Secondary School Arithmetic Test (see Figure 8-2).

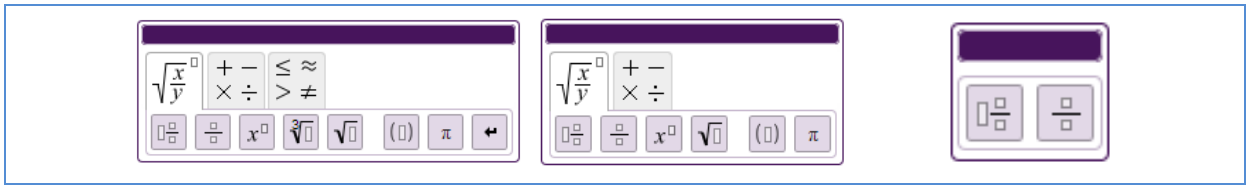


Figure 8-2. The formula editor in full, limited, or minimal versions

Drawing graphs.

The graphic module of the Digital Mathematics Environment (Utrecht University) was made suitable for use in the QuestifyBuilder - Facet chain (see the upper left in Figure 8-1). This makes it possible to design tasks in which students can, for example, draw points and charts. The “Custom Interactions” feature can be used to embed these activities in tasks and make the student’s response available for automated assessment.

A sample task is displayed in the left screen of Figure 8-3. The question involves using two points to draw the sum graph of the two graphs that are given. In the right screen, the student has clicked two points and drawn a line through them. The automated assessment system ensures that any pair of points along that line is recognized as correct. This provides students with flexibility and makes it possible to recognize different manners of solving problems as correct.

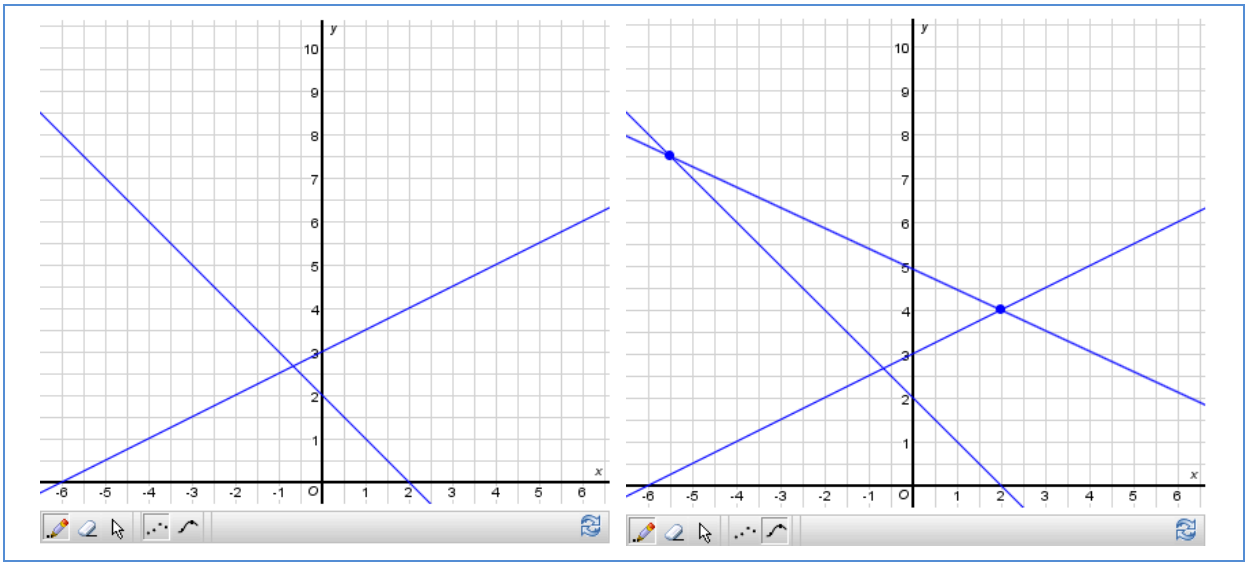


Figure 8-3. A hypothetical task using the graphic model of the Digital Mathematics Environment (DME)

Performing geometric constructions.

The embedding of a dynamic-geometry package, GeoGebra, makes it possible for students to create geometric constructions in Facet (see the upper right in Figure 8-1). For example, consider the construction of circles or altitudes. These constructions are also suitable for automated assessment. GeoGebra is a relatively extensive package, which offers more options than are needed for the DET in many cases. As with the formula editor, the menu bar can be restricted. Depending on the task, the task constructor can select the buttons that will be available to the student. Two variants are displayed in Figure 8-4. It is obviously advisable to aim for a limited number of menu-bar variants, so that students can become accustomed to them.

Figure 8-4. Two variants of the GeoGebra menu bar

The second key point concerns the ***"intelligence" of automated assessment***. In this context, the term "intelligence" refers to the extent to which tasks in which the construction space allows candidates to provide correct answers in many different ways are assessed accordingly. The assessment is thus intelligent to the extent that it recognizes all correct responses as such. This is realized in two different ways: One way involves using Boolean variables in GeoGebra (see Box 8-1), and the other using a computer algebra system (CAS), in this case, the open source program Maxima (see the lower right of Figure 8-1). The deployed computer algebra system can check, among other things, whether an expression is exactly equal or algebraically equivalent to the key, and whether an algebraic expression meets a specified criterion (see Box 8-1). To set this assessment method (with the exception of the latter), the test constructor can use the appropriate options in the QuestifyBuilder scoring module, shown in Figure 8-5 (right).

Box 8-1. Realized automated evaluation

- The *equal strict* setting instructs the CAS to determine whether an expression is an exact match to the key entered by the test constructor. This setting thus results is a very strict assessment: if the key has been entered as x+3, the response 3+x will be assessed as incorrect.
- The *equivalent* setting instructs the CAS to determine whether the key and the response are algebraically equivalent. This assessment is not as strict. If the key has been entered as x+3, the response 3+x will also be assessed as correct, as will the responses (x^2-9)/(x-3) and x+1+1+1.
- The *equal soft* setting, which is not a standard option in Maxima, but for which a script has been developed, maintains a strictness level midway between the two aforementioned options. If the key has been entered as x+3, the response 3+x will be assessed as correct, but the less simplified forms (x^2-9)/(x-3) en x+1+1+1 will not.
- The *evaluate* setting instructs the CAS to determine whether an algebraic expression meets a criterion specified by the task constructor. This option is particularly useful for tasks that call for generating expressions that meet certain conditions. For example, suppose that the question involves using an equation to draw a parabola running through the point (3, 9). In principle, the number of parabolas to which this applies is infinite, as are the corresponding formulas. To assess the response, the stated expression is evaluated to determine whether it is indeed quadratic and whether the value for x=3 is equal to 9.
- The *dependency* setting relates to several different numerical entry fields or different point coordinates. Responses are evaluated to determine whether the second entry is related to the first in a particular way. One use of this option is to determine whether an incorrect answer has been calculated further in the correct manner. A second application involves determining whether a point in the space meets certain criteria. In the task depicted in Figure 8-3, a reading is made of the coordinates of the two points entered by the student. The first coordinate of the first point (*a*) may take any value. The second coordinate, however, must ensure that the point is located along the requested line with the equation y = 5 - x/2. To this end, the task constructor uses the dependency setting to enter the condition b = 5 - a/2. For the coordinates of the second point, (c, d), this specifies the supplementary dependencies that c is not equal to a and that d = 5 - c/2.
- The *equal equation* setting instructs the CAS to determine whether two successive equations are equivalent (i.e., whether they have the same solution compositions).
- The use of Boolean variables is a specific feature of GeoGebra. A Boolean variable is defined by a test constructor in GeoGebra and has a value of either 0 or 1. For example, if the candidate must define a point A on the line with the equation x = 2y + 13 and with the x- coordinate not being equal to 3, the Boolean variable in Figure 8-5 (left) can determine whether the candidate has performed this correctly. The value of this variable, 0 (false) or 1 (true), appears in Facet and can be assessed automatically.

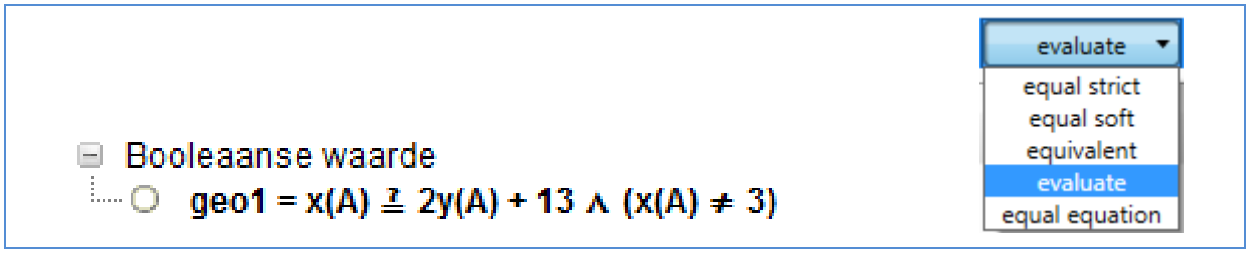


Figure 8-5. Tools for the test constructor: Boolean variable in GeoGebra (left) and options menu in Questify (right)

The third key point concerns the ***assessment of intermediate steps***. This is the trickiest of the three points, and it has been realized within the framework of the DET only to a limited extent. One way to do this involves interpreting incorrect alternatives in multiple-choice tasks or incorrect answers in open-ended tasks in terms of an error in an intermediate step. If a task calls for calculating 3*4^2, the answer 144 (or the selection of 144 as an alternative) indicates that the squaring was performed after the multiplication, and not before. This approach is not used very frequently in the DET. A second manner involves explicitly asking students to make notes of their intermediate steps and using the dependency option. For example, if an task asks how much money will be in a savings account after two years, with the first year having an interest rate of 3% and the second year having an interest rate of 2%, an intermediate step could consist of

asking what the amount would be after one year. If this answer is not correct, the dependency option makes it possible to determine whether the final answer is nevertheless 1.02 times greater than the intermediate answer. One disadvantage of this method is that the partial steps must be carefully structured in advance, even though conceiving of a sequence of intermediate steps is a characteristic of a candidate's problem-solving ability. This method was therefore not used very much. A third approach, which has not yet been realized, involves the use of multi-line answer fields and "domain reasoners" to assess the intermediate steps of students (Drijvers, submitted). This approach is the subject of the Erasmus+ project Advise-Me.

8.4 Future prospects

Important progress has been achieved within the DET with regard to the automated assessment of mathematics. It would be worthwhile to apply the DET results to other digital arithmetic and mathematics tests in which automated assessment plays a role. Examples include the Secondary School Arithmetic Test and the computer examinations for vmbo, as well as the Adaptive Central Final Test. Third, the possibilities require further development, as is currently taking place within the Erasmus+ project Advise-Me, in which "domain reasoners" and student modeling are being used to develop the assessment of intermediate steps (see http://advise-me.ou.nl).

8.5 References

Drijvers, P. (submitted). Digital assessment of mathematics: opportunities, issues and criteria. *Mesure et évaluation en education*.

Drijvers, P., & van Reeuwijk, M. (2015). *Automatische beoordeling van wiskunde [Automated assessment of mathematics].* Report of the expert meeting in Arnhem, February 16-18, 2015. Arnhem/Utrecht: Cito/CvTE.

Drijvers, P., Visser, S., Straat, H., & Schouwstra, S. (2014). Digitale diagnostische toetsing van wiskunde. Onderzoek van een nieuwe omgeving [Digital diagnostic assessment for mathematics: Research on a new environment]. In S. Schouwstra (Ed.). *Diagnostische tussentijdse toets: Onderzoek 2014 [Diagnostic Educational Test: Research 2014]* (pp. 53-70). Arnhem: Cito.

Fife, J. H. (2011). *Automated scoring of CBAL mathematics tasks with m-rater. Research Memorandum*. Princeton, NJ: ETS. http://www.ets.org/Media/Research/pdf/RM-11-12.pdf.

Groenheiden, M. (2016). *Intuïtieve Interacties van leerlingen met digitale hulpmiddelen tijdens de Diagnostische Tussentijdse Toets Wiskunde. Intern rapport [Intuitive interactions of students with digital tools during the Diagnostic Educational Test in Mathematics: Internal report].* Arnhem: Cito.

Stacey, K. & Wiliam, D. (2013). Technology and Assessment in Mathematics. In M. A. Clements, A. Bishop, C. Keitel, J. Kilpatrick, & F. Leung (Eds.), *Third International Handbook of Mathematics Educatio*n (pp. 721-751). New York/Berlin: Springer.

Williamson, D. M., Mislevy, R. J., & Bejar, I. (Eds.) (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

9 Exchange of information between the authoring environment and Facet

Arjan Aarnink

9.1 Introduction

Test-and-exam chains often involve multiple stakeholders, multiple suppliers, and multiple systems. To ensure that the exchange of information and the links to systems would proceed as smoothly as possible, open standards were used. Open standards are well documented, unrestricted, and available to use free of charge. In the Facet chain (of which the DET was part) the Question & Test Interoperability standard (QTI; https://www.imsglobal.org/activity/qtiapip). This standard is managed by IMS Global Learning Consortium (https://www.imsglobal.org/).

It is intended for the exchange of test materials, with QTI describing the test, items, layout, scoring, and candidates' results. In this chapter the terms *item* and *interaction* are used in the way they are used in the QTI standard (see also Chapter 3). An overall description of the QTI features is provided in Figure 9-1 below (Source: IMS site: http://www.imsglobal.org/activity/qtiapip).

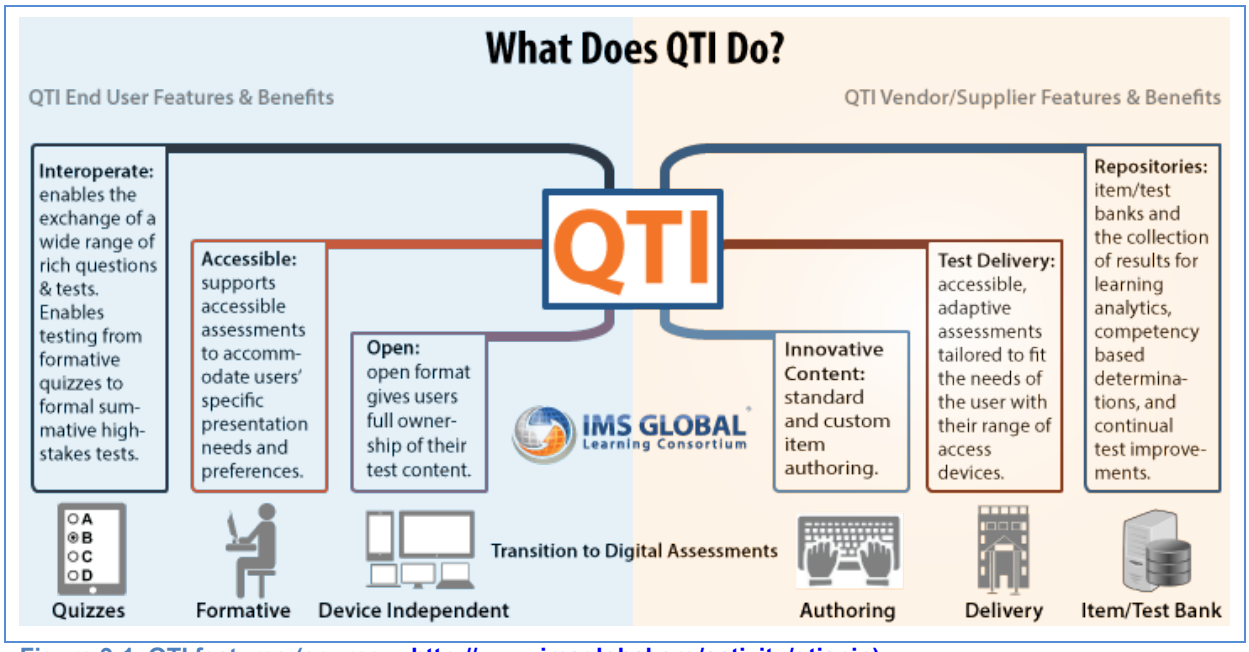


Figure 9-1. QTI features (source: : http://www.imsglobal.org/activity/qtiapip)

9.2 Dutch Exam Profile

The exact use of the QTI standard for Facet is described in the "Dutch Exam Profile" (DEP). This application profile includes extensions to the international standard for the exchange of IMS QTI test materials (see Edustandaard, 2017). Version 2.1 was in use at the time of the DET project.

The DEP extensions were added in order to support functionalities that were not yet supported by QTI (e.g., special behavior and presentation of items, see section 9.4). The DEP also allows for the description of special forms of scoring. The goal is to restrict these extensions to a minimum. Cito is also working to support the relevant functionalities in new versions of QTI. Many functions that are yet specified in DEP will become available in QTI when versions QTI2.2 and QTI3 are released. Plans call for largely dismantling the DEP in the future, thereby enhancing the inter-operability of test materials.

The DEP is managed by Edustandaard, the organization devoted to educational standards within the Dutch education system. Additional information on the DEP standard is available on the Edustandaard website (https://www.edustandard.nl/standard_afspraken/dutch-exam-profile-dep/).

9.3 Application of standards within the chain

Roughly, the QTI standard is used at two points in the Facet chain:

1) *Exchange between the Cito authoring system (Questify Builder) and Facet*
   This exchange is depicted in the green block to the left of the scheme in Figure 9-2. It consists of a DEP package containing all information about the items and tests. The DEP package is offered as a ZIP file, which is signed and encrypted, so that the content can be read only by the parties who are authorized to do so. A Facet manifest is included, containing a description of the parties for whom the test is intended. For example, the descriptions include the target group (e.g., educational stream) and administration period (the period within which the test can be administered).

2) *Exchange of student results between Facet and Cito analysis tools*
   This exchange is depicted in the green block to the right of the scheme in Figure 9-2. In this step, student answers and outcomes are exchanged between Facet and Cito. This exchange, which is referred to as "data dump" in the Facet chain, is an XML exchange based on the QTI scheme for "Result Reporting." Additional information is available on the IMS website (http://www.imsglobal.org).

9.4 Necessary exceptions and extensions

The DET is an innovative test. Although the goal was to align as closely as possible to the QTI standard, this was not possible on a number of points. An overview of adjustments that were applied for the purposes of the DET is provided below:

Use of the existing QTI expansion possibilities

In a few cases, the DET called for exceptional behavior on the part of existing types of tasks. The QTI standard offers the possibility to force different behavior. The following are several examples.

- An existing QTI task type known as "hot-text interaction" (marking task) was used to divide a text fragment into paragraphs. *Class names* were used to create exceptions to the standard behavior. In a marking task the student can select parts of a text that are marked. In addition, a paragraph break appears above the marked sentence in a paragraph task.
- An existing QTI task type known as "graphic GapMatch interaction" (dragging task) was used to divide images into categories. *Class names* were used to create exceptions to the standard behavior. In a drag task with image ("graphic GapMatch interaction") an element (answer option) can be dragged to a place on an image. In a categorization task exercise, a student can drag multiple elements to a specific place on an image.

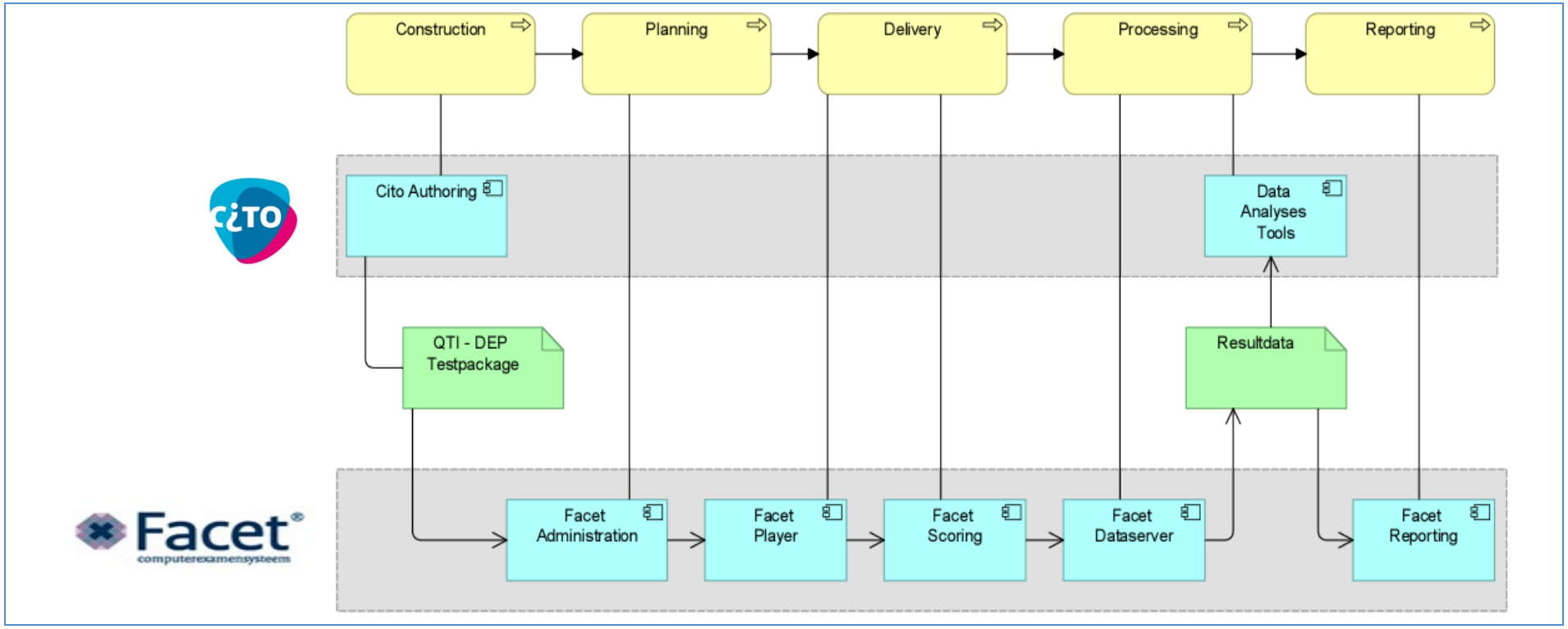


Figure 9-2. Representation of the information exchange between the authoring environment and Facet

Extensions to QTI

The QTI standard did not offer space for a few types of items that were necessary for the DET. Extensions were added to QTI so that such items could be offered. The standard manner of extension existing within the Facet chain was used to this end. This was accomplished by including the extensions in DEP. The following are several examples.

- An extension was created in order to allow the adaptive administration of the DET. This extension offered the possibility of providing an adaptive module and driver along with the test.
- Another extension was created for items calling for the inclusion of mathematical symbols and images (e.g., in “inline choice” interactions or drop-down task).
- It was not possible to offer a formula editor as a default. An extension made this possible. In addition, item authors could specify the buttons of the formula editor that would be available to candidates.

Custom Interactions

Advanced interactions could be offered using the custom interactions (CI) of QTI. Custom interactions are a semi-structured manner of offering candidates interactions based on Java script and HTML5. These CI are provided by Cito. An agreement was reached between Cito and the Executive Agency for the Department of Education [DUO] to determine how the CI results should be stored in Facet.

Some examples:

- Interactions based on the DME (Digital Mathematics Environment)
    - Additional information is available on the website of the Freudenthal Institute (http://www.fi.uu.nl/wisweb/en/)
- Interactions based on Geogebra
    - Additional information is available on the Geogebra website (https://www.geogebra.org/)

Custom operators

Facet uses a computer algebra system (CAS) to ensure the proper automated assessment of mathematical items. The QTI package includes Cito scoring rules for Facet concerning how the relevant mathematical items should be scored. These rules are packaged in “custom operators,” which are a structured manner of including information in QTI.

Some examples:

- Custom operator for determining whether two formulas are equal.
- Custom operator for determining whether two formulas are equivalent.

9.5 Future developments

The preceding sections describe the points for which exceptions were made to the QTI standard for the DET. This section addresses developments that have occurred with regard to standards since the development of the DET.

Extensions to QTI

At the launch of the DET, QTI Version 2.1 was available. It was necessary to apply extensions to QTI (through the DEP) in order to make essential functionalities available for the DET. Since that time, QTI2.2 has been released. This version provides many functionalities that make a great many DEP extensions unnecessary. Examples include the support of mathML for the presentation of mathematical formulas.

Nearly all DEP extensions will become obsolete when QTI3 is released (expected in 2018). Cito is playing an active role in the development of QTI, and it has contributed knowledge (including from the DET) to the development of QTI3.

Custom Interactions

Custom interactions (CI) are very powerful. Guaranteeing proper operations nevertheless required a bilateral agreement between Cito and the Executive Agency for the Department of Education [DUO]. The arrival of an IMS standard for CI (expected in 2018) will eliminate the necessity of bilateral agreements and allow for CI inter-operability.

Computer Adaptive Testing (CAT)

Cito possesses considerable knowledge with regard to the integration of adaptive logic within a QTI environment. This knowledge is shared with IMS. In 2016, Cito and several other IMS members initiated a working group for the development of a CAT standard. This standard was introduced in late 2017. Finalization is expected in 2018. This will largely eliminate the need to make bilateral agreements between suppliers of CAT modules and administrative environments.

9.6 References

Edustandaard. (2017). *Dutch Exam Profile (DEP).* Retrieved november 2017, from https://www.edustandaard.nl/standaard_afspraken/dutch-exam-profile-dep/dutch-exam-profile-versie-4-0/.

IMS Global Learning Consortium (2017). *Question and Test Interoperability and Accessible Portable Item Protocol Background.* Retrieved november 2017, from https://www.imsglobal.org/activity/qtiapip.

10 Data processing and item analyses

Jesse Koops

10.1 Introduction

The regular ICT chain of development, administration, data storage and analysis for the DET was complicated (see the previous Chapter 9). In addition, a large share of this chain was either not available or still missing important functionalities during the development of the DET between 2012 and 2017. Another issue was that the intended changes or additional functionalities were to be included in the standard process of test development and administration, and this is time-consuming. At the same time, the results of the changes and additional functionalities for the DET were to be immediately available for the analyses and further development of the DET. For example, this was the case for the recoding of key errors, the linkage of tasks and responses to the substantive aspects and the automated assessment of mathematical formulas. These functionalities were to be immediately available for the DET analyses, even though they had not yet been implemented in the standard process.

For this reason, we worked with a system that was developed especially for the pre-test phase and that gradually grew during the development of the DET. It consisted of a relational database and some for the DET adjusted tools and libraries (i.e., collections of functions that can be called up by programs) that had been developed more or less especially for the DET. These facilities could be quite easily adjusted from year to year, thus allowing them to meet the continuously changing needs and software landscape of the DET. One disadvantage of this approach was that the changes that have been implemented each year (resulting in an altered data structure) make it difficult to combine data from different years for a single analysis.

The following is a list of tools that were at least partly developed or adjusted for the DET:

- Relational databases (differing by year)
- Scripts for reading QTI items (tasks, see previous Chapter), Facet data and other data sources
- Scripts for exporting data from the database to the format that is needed for calibration and block structuring (see also Chapter 12)
- QTI scoring engine (2015-2017)
- Temporary scoring systems (2013-2015)
- Key-correction tools (2013-2015)
- Computer algebra system (CAS) for scoring open-ended mathematical answers, based on SymPy (www.sympy.org/en/index.html).
- Online viewer for QTI items (tasks)
- Dashboards and other data viewers
- Online test and item-analysis applications
- 3DC standard-setting application (Cito, 2018)

At this point, it is important to note that, although the development of these tools was necessary to the further development of the DET, they are not a part of the end product delivered. For several reasons, delivery would also not be useful. Several tools have since become outdated by the software chain, and others have been specifically adjusted to the single-use data structure for a given year or for the systems used by Cito.

Whenever possible, Cito generalizes and standardizes internal DET applications, so that it can be used by everyone and benefit the educational system. Of the items in the list above, the 3DC standard-setting application was published as free software (Cito, 2018). Furthermore, the online test and item-analysis[9]

[9] Although it actually entails item-response analysis, it is usually referred to as item analysis. For this reason, we have adopted the term *item analysis* in this section.

applications inspired a graphic user interface for the existing Dexter test and item-analysis package (Maris, Bechger, Koops en Partchev, 2017).

10.2 Data processing

After the data were received, they are read from the XML files delivered by the Executive Agency for the Department of Education [DUO], based on the item and test definitions in the Questify packages. The data arrived in a relational database. They were subsequently inspected for completeness based on lists of participating schools and subjects.

Once the data were complete, several automatic inspections were performed on the keys and scoring, followed by an initial quick exploratory analysis for detecting structural problems that could indicate technical errors in the administration, scoring or reading of the files. The data-processing flow chart for 2016 is presented below as an illustration (Figure 10-1).

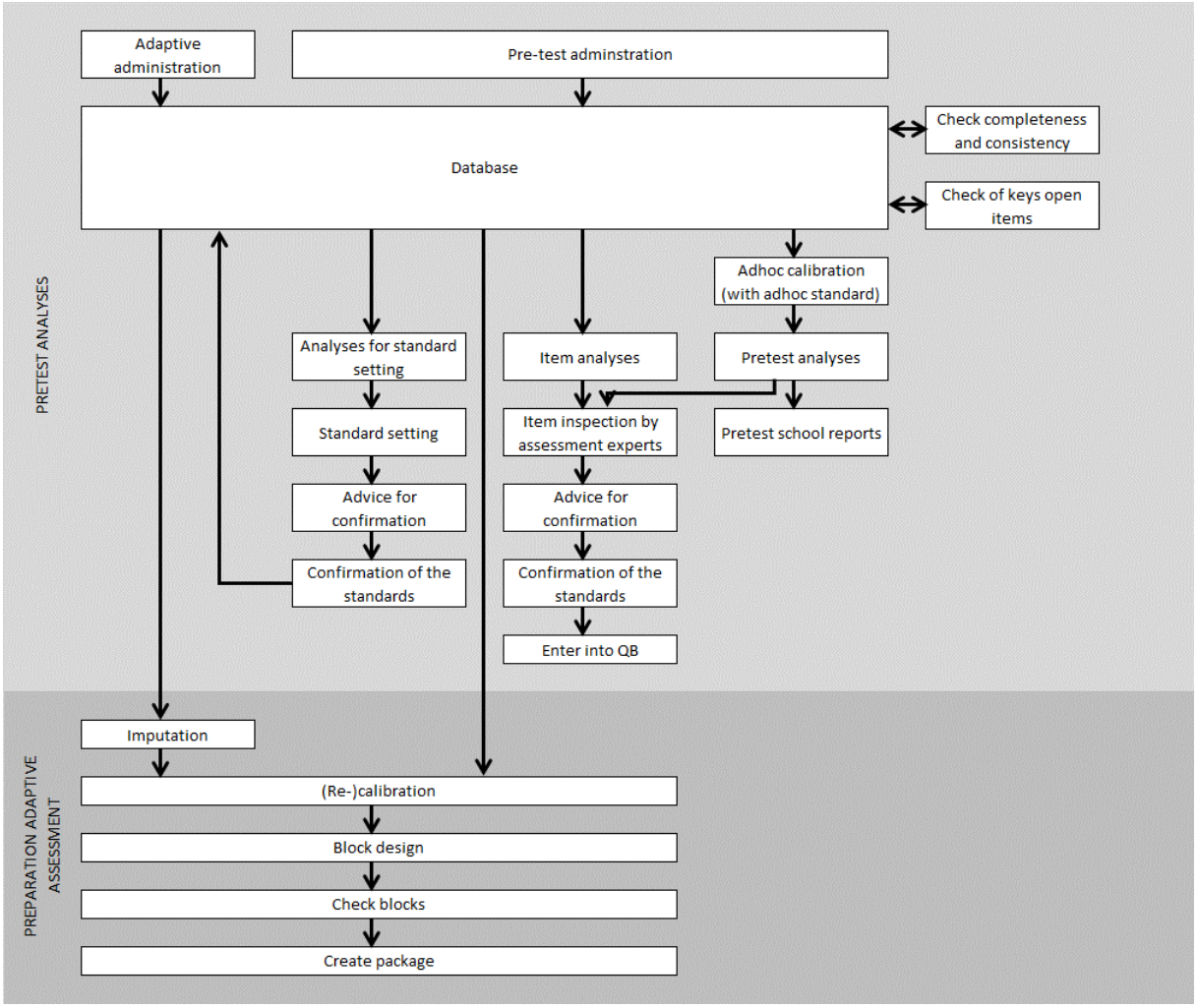


Figure 10-1. The DET data-processing flow chart for 2016 (from Cito, 2016)

10.2.1 Data processing by year

The following is a brief description of the most important data-processing elements from each year in the development of the DET.

2013

In the first year of the development of the DET, we still had to use the old ExamenTester administration environment of Cito. ExamenTester was developed for use with examinations, and it is thus not specifically designed for diagnostic tests. For example, the DET calls for more complex types of tasks, scoring and metadata than are possible in ExamenTester. For the initial try-out, this meant that only 45% of all tasks could be tested automatically using ExamenTester.

The following measures were taken in order to ensure that the other tasks could be processed in the manner required by the DET:

- The assessment experts compiled assessment instructions for the assessment of open-ended answers that could not be scored automatically.
- For the manual assessment of open-ended answers, assessors were recruited to assess all open-ended responses within a period of four weeks, including the writing assignments for the languages. In all, the assessors spent more than 850 hours on the manual assessment of the open-ended responses.
- Two temporary assessment tools were developed for the assessment of all non-numerical open-ended tasks.
    - One tool for the open-ended text tasks
    - One tool for the mathematics assignments that were created with the mathematical toolbox (Flash)
- All tasks with multiple responses and all 'closed-ended' tasks with Flash content[10] were assessed automatically through a database. Specific routines were developed for this purpose.

2014

In 2014, Questify Builder was used for the first time as an authoring environment, with Facet used as an administration environment. In theory, all tasks could be scored automatically, with the exception of the essay assignments for Dutch and English. The assessment was still performed outside the Questify-Facet chain, as Questify Builder was not yet capable of providing keys for all possible types of tasks, and it was not yet possible to score all types of tasks automatically in Facet. Another new development in 2014 involved a few tasks in which a formula editor could be used. In this year, cautious experiments were conducted with the automated scoring of formula answers.

At that time, Questify and Facet offered no facilities for coding tasks with multiple interactions. The coding of separate responses was performed in a database. Assessment experts were able to correct keys with the key-correction tool that had been developed for the DET.

2015

The further development of Facet and Questify ensured that, in 2015, tasks could be coded at the level of interactions. Because it was not yet possible to score all tasks automatically, and because re-scoring was not possible in Facet (which is essential for a pre-test with open-ended tasks), the automated scoring and coding features in Facet were not used. One major step forward involved the ability to read the codings automatically at the interaction level from the QTI item definitions exported from Questify Builder.

For scoring answers to short open-ended tasks with formula entry, a modest library based on the open-source library SymPy was developed internally in 2015.

2016

The conditions for data processing in 2016 were comparable to those of 2015. In 2016, it was possible to work with a more advanced database system (PostgreSQL instead of SQLite), as well as with an internal server on which online applications could be developed and installed. This made the item analysis and key correction noticeably smoother (see the following section for a description).

[10] In ExamenTester, newer forms of tasks could be created only with Adobe Flash, a computer program that can be used to create animations, web videos and web applications.

2017

The Questify-Facet chain was practically complete by this time. The data processing was comparable to that of 2016. It was possible to automatically evaluate short open-ended tasks with formula entry in Facet. Re-scoring was also possible in Facet, albeit with several limitations. Because the internally developed system was not subject to the same limitations as Facet, the internally developed evaluation module was used for most purposes. Facet evaluation (scoring) was used for purposes of verification.

10.2.2 Learning points in the DET chain

The tools that were developed on an *ad hoc* basis were used to circumvent certain problems in the DET chain. Some of these problems were of a temporary nature and inherent to the rapid development of the project and the stage it was in at that time. A few other problems remained at play until the end of the project. In retrospect, we were able to identify two general causes for the problems that were experienced (in addition to time pressure and other circumstances).

Modularity

In software, modularity refers to the extent to which 'modules' are separated from each other. A module is a self-standing part of a program or library that is specialized in one specific task or a set of closely related tasks. In this way, a module can be used to carry out the same tasks in different applications. A high extent of modularity ensures that an adjustment in one module does not lead to adjustments in other modules. In addition, modules can be used separately from each other.

An example in which sufficient modularity has not been achieved within the DET software chain has to do with the re-coding (re-scoring) of student answers. A modular solution (which is used at Cito to work around the limitations of the DET chain) can consist of a module that expects input in the form of the evaluation rule and a student answer, and that can use this input to generate an evaluation for the task. One such module can be used in multiple places: for scoring in a testing system, as well as for retrospective scoring with the same or an adjusted key.

The evaluation and revision of the evaluation (re-scoring) in Facet could not be disconnected from the administration software, and the ultimate implementation requires students to proceed virtually through the entire test once again. For this reason, re-scoring has since been enabled in Facet, but not for an adaptive administration (as this would influence the testing path). One consequence is that the re-scoring of seed tasks in the adaptive administration was also carried out at Cito (instead of through Facet) in the last year of development. In a modular solution, the actual progress through the test would be functionally separated from the scoring of the tasks.

Data integrity and identification variables

Variables that are used to identify data elements ('identifiers', e.g., item codes or test IDs) must meet at least two conditions:

1. *Unique* for an element within the scope in which they are used (i.e., the same identification code is not used for different elements, and the same element always has the same identification code)
2. *Stable* (as long as an element does not change, the identification code also does not change)

There was much room for improvement, particularly with regard to stability. For example, different identifications were used for the concept codings in different years and in different systems; identifiers for response options changed during updates of Questify; and the identification codes of the tasks (task IDs) were not consistent with regard to the use of upper-case and lower-case letters and minus signs or underscores. In addition, it was not possible to rule out the possibility of 'freezing' – adjusting tasks in the authoring environment (Questify Builder) – and, in many cases, it was not possible to determine whether an task had been adjusted and by whom.

10.3 Item analysis and key correction

To allow for the evaluation of the quality of tasks, descriptive item analyses were performed to provide information on the statistical (and other) properties of the tasks that have been developed.

For task evaluation and key correction, the possibility for assessment experts to evaluate a task in the context of the data and psychometric information is particularly important. In the standard DET chain, the task and the key were visible in the authoring environment (Questify Builder), the scoring was performed in Facet, and the response data were available in a Microsoft Server database. Psychometric data are usually the result of analysis with software (e.g., SPSS or R and, in the case of the DET, a Fortran routine). Analytic results are usually delivered as text, CSF, or Excel files, or as images.

Such fragmentation of the information that assessment experts need in order to evaluate a task is not helpful. For this reason, an online application was developed internally, in which these sources of information were combined as much as possible.

To allow for the evaluation of the quality of tasks, descriptive task analyses were performed to provide information on the statistical (and other) properties of the tasks that had been developed. For each subject, there were different versions of the test with multiple sections (one section for each main aspect) within each educational stream, so that all tasks could be pre-tested. For the most important psychometric measures that were used for the assessment of the quality of tasks, see Table 10-1.

Table 10-1. Description of the most important psychometric measures for evaluating the quality of tasks

Measure	Description
P-value	P-value (proportion correct) of the concept response. If the response is polytomous, the mean would be displayed here.
Rit value, Rir value	Correlation of the concept response with the summed score of the test and the correlation of the concept response with the test, less the relevant concept response. The Rit and Rir values are measures of the discriminating power of tasks. In the case of an adaptive test, the correlation with the block in which the task appeared is used, although it is only informative in the first blocks (in Chapters 12 and 13 the block design is discussed) that were presented to all students.
Cronbach's alpha, Cronbach's alpha - task	Cronbach's alpha is a lower boundary for the reliability of the test score. For this measure as well, it can be interpreted only in the pre-test or in the first blocks in the adaptive test.
$\mathbf{P_{below}}$	The likelihood that students who are below level (20% of the pre-test students) will answer the concept response correctly, after *ad hoc* calibration.
$\mathbf{P_{at}}$	The likelihood that students who are at level (60% of the pre-test students) will answer the concept response correctly, after *ad hoc* calibration.
$\mathbf{P_{above}}$	The likelihood that students who are above level (20% of the pre-test students) will answer the concept response correctly, after *ad hoc* calibration.

The assessment experts could see the outcomes of the test and task analyses in the online application (see Figure 10-2). The assessment experts could write comments for each task. If desired, each task could be displayed (see Figure 10-3), and an Excel file containing the outcomes could be downloaded.

All of the tasks that could be problematic were marked in an Excel file, and they were subjected to substantive inspection by the assessment experts and discussed with the confirmation committee (see also Chapter 2). The confirmation committee assessed the tasks that had been submitted and determined whether to approve or reject the tasks that they had discussed.

Figure 10-2. Screen print of the test and item analyses

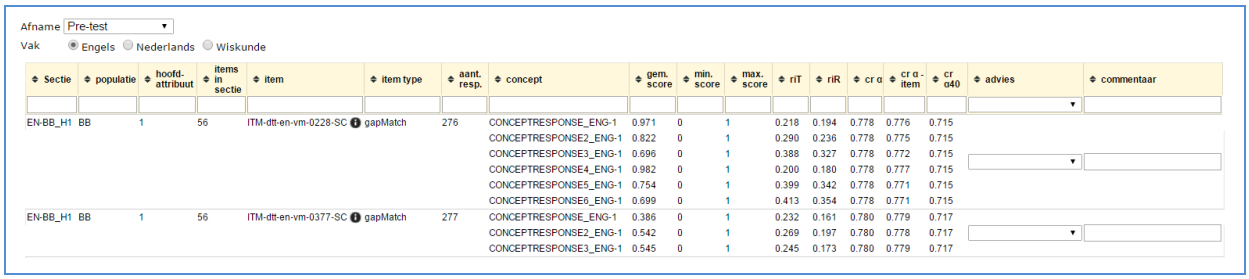
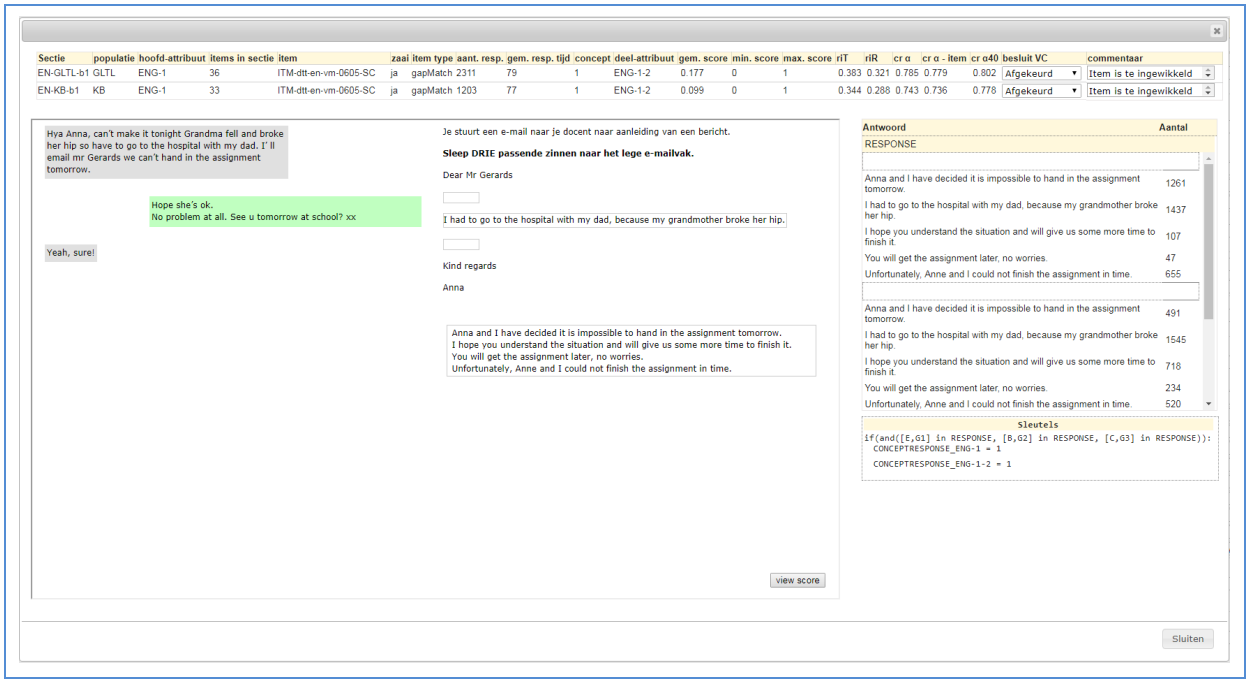


Figure 10-3. Screen print of the test and item analyses, zooming in on a task.



10.4 References

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016 [Diagnostic Educational Test: Pre-test report 2016].* Arnhem: Cito.

Cito (2018). *Computerized standardsetting using 3DC.* Retrieved from http://www.cito.com/our-expertise/implementation/3dc

Maris, G., Bechger, T., Koops, J. & Partchev, I. (2017). *Dexter: Data Management and Analysis of Tests*. R package version 0.6.1.

R Core Team (2013*). R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

11 Standard setting for the DET

Sanneke Schouwstra, Jesse Koops, and Karen Keune

11.1 Objective of the standard setting

In the DET, diagnoses are made with regard to Dutch writing skills, English writing skills and mathematics. In this process, the category of mastery into which the student falls is determined for each aspect of the student model (knowledge aspects or sub-skills): below, at, or above level. The objective of the standard setting was to determine the boundaries between below level and at level and between at level and above level for each main aspect and for each educational stream. To this end, standards were set for each subject and in each educational stream (vmbo-bb, vmbo-kb, vmbo-gt, havo and vwo) for each of the main aspects. In all, 15 standard-setting procedures were conducted with experts following each pre-test. This standard-setting procedure was performed using the the Data-Driven Direct Consensus (3DC) method (Feskens, Keuning, Van Til, & Verheyen, 2014; Keuning, Straat, & Feskens, 2017).

11.2 The 3DC method

The 3DC method (Feskens, Keuning, Van Til, & Verheyen, 2014) is relatively new, and it combines elements from the Angoff procedure (Angoff, 1971), the Bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001), and the Direct Consensus Method (Sireci, Hambleton, & Pitoniak, 2004). In 2013, when the overall plan for the standard-setting procedure for the DET was elaborated (Keune & Schouwstra, 2014), this method had already been applied successfully in the standard-setting procedure for the tests MBO-COE Dutch and MBO-COE Arithmetic, as well as for setting the performance standards for the Common European Framework of Reference for Languages (Feskens, Keuning, Van Til, & Verheyen, 2014). The 3DC method is based on the notion that the tasks in a test can be subdivided into clusters. This method is therefore particularly well suited for tests containing clearly distinguishable clusters of tasks, as is the case with the DET. The tasks in the DET are clustered according to aspects of the student model. For each cluster of tasks, it can then be determined how many responses must be answered correctly in order to be classed in the intended category of mastery.

The 3 DC standard-setting procedure is performed with a team of experts in the subject area. With the goal of arriving at well-founded standards, the experts were provided with as many different sources of information as possible. Prior to the standard-setting session of the DET, the experts received the interim objectives developed by Netherlands Institute of Curriculum Development (SLO, 2012) and the Assessment Specification (College voor Toetsen en Examens, 2014). The SLO interim objectives include descriptions of what students should be able to do and know halfway through the lower years of secondary education. The Assessment Specification prepared by the Board of Tests and Examinations (CvTE) describes the aspects that should be diagnosed. During the session, the clusters of tasks are presented to the experts, who draw upon their own experience as teachers in their judgements. During the standard-setting session, the experts also receive empirical feedback on the relative difficulty of the tasks, on the judgements of the other experts and, possibly, on the impact of the judgements on the number of students classed in a given level (below level, at level, or above level).

11.3 The experts

The goal of the Board of Tests and Examinations was to recruit 10-12 experts for each standard-setting procedure. On average, nine teachers ultimately participated in each session (see Table 11-1). The experts were recruited primarily from the pilot schools, given the importance of ensuring that the substantive experts had recent experience with teaching the target group.

In 2015, 93 teachers participated (of the 107 teachers who had signed up), and 82 participated in 2016 (of the 90 teachers who had signed up). Half of the teachers signing up in 2016 had already participated in the standard-setting procedure in 2015. To ensure consistency in the judgements, the boundaries for the various educational streams were determined in part by the same group of experts. In 2015, 40% of the

teachers participating had been involved in multiple standard-setting procedures (e.g., for havo and for vwo). In 2016, the share was 47% (see Table 11-2).

Table 11-1. Number of experts for each standard-setting procedure

		Number of experts for each standard-setting procedure				
		vmbo-bb	vmbo-kb	vmbo-gt	havo	vwo
2015	English Writing Skills	8	10	9	12	11
	Dutch Writing Skills	7	9	7	8	10
	Mathematics	8	10	11	8	10
2016	English Writing Skills	10	11	10	8	7
	Dutch Writing Skills	7	8	12	12	11
	Mathematics	5	8	9	11	12

Table 11-2. Number of experts participating in the standard-setting procedures

		Signed up	Participated	Multiple sessions
2015	English Writing Skills	34	32	50%
	Dutch Writing Skills	36	29	38%
	Mathematics	39	34	32%
2016	English Writing Skills	27	22	77%
	Dutch Writing Skills	35	33	36%
	Mathematics	32	30	37%

11.4 The tasks

The 3DC is a test-oriented standard-setting method, as it involves the experts basing their judgements concerning the boundaries on the content of the test. For the DET the experts were expected to arrive at an judgement concerning the number of responses that a borderline student should answer correctly for each cluster of tasks. In each of the two pre-tests, however, the total number of tasks administered[11] in each educational stream exceeded the number that the experts were able to review in one day, for this reason a selection of tasks was used.

In 2015, ten to twelve tasks were selected for each cluster. In that year, the standard-setting procedure focused on the diagnoses that would be reported in the first, limited adaptive administration in 2016. In the limited adaptive administration, the reports on the languages were to be based exclusively on the main aspects, while the reports on mathematics were to include all sub-aspects of only two domains. For the languages in 2015, therefore, boundaries were determined between below level and at level and between at level and above level for four clusters (one cluster for each main aspect), each consisting of about twelve tasks. For mathematics, boundaries were set for six clusters (for each sub-aspect of Domain D "*Measurement and geometry*" and Domain E "*Relationships and formulas*") for each cluster of approximately 10 tasks.

In 2016, the standard-setting procedure focused on the diagnoses of the fully adaptive administration in 2017. In 2016, boundaries were set for each sub-aspect, as was the case for mathematics. These boundaries were subsequently combined into one standard for each main aspect. Eight clusters were included for English, with Dutch ranging from twelve (vmbo) to thirteen (havo/vwo) and mathematics ranging

[11] In order to pre-test all of the items that had been developed, a design was used in which each student was required to complete only a part of the items (Cito, 2015; Cito, 2016).

from six (vmbo) to nine (havo/vwo) clusters. Given the necessity of judging more clusters for the languages than had been the case in 2015, a guideline was established that at least twelve responses were to be included for each cluster (instead of the twelve tasks in 2015). On average, there were fourteen responses in each cluster for the languages, with an average of sixteen responses in each cluster for mathematics (as had been the case in 2015). The average number of tasks and responses used for the standard-setting procedures is displayed in Table 11-3.

To include a greater variety of task types, contexts and/or genres in the standard-setting procedure, maximum use was made of tasks with a limited number of responses. Complex[12] and problematic[13] tasks were not used. If too many tasks remained after the complex and problematic tasks had been excluded, tasks that were also used in the adjacent educational streams were selected first, followed by a random selection of remaining options. The use of tasks that are administered in adjacent educational streams makes it possible to examine whether the standards in the adjacent educational streams are consistent with each other (e.g., whether higher requirements are set for vmbo-bb than is the case for vmbo-kb).

Table 11-3. Average number of tasks and responses in the standard-setting procedures

		English Writing Skills		Dutch Writing Skills		Mathematics	
		Tasks	Responses	Tasks	Responses	Tasks	Responses
2015	Average per cluster	11	23	13	54	10	16
2016	Average per cluster	11	14	5	14	10	16

11.5 The task of the experts

The experts received information by post concerning the standard-setting procedure, the interim objectives, and the Assessment Specification. Prior to the standard-setting procedure, the experts received a short training session providing details about the DET, the standard-setting procedure, and the computer applications that would be used during the standard-setting procedure. The 15 standard-setting procedures were conducted within a period of two weeks, after all data from the pre-test had been processed, the test and task analyses had been conducted, and any key corrections had been implemented. The duration of each standard-setting procedure was one day. Each day, no more than two standard-setting procedures for two different subjects were held simultaneously, such that the experts would be able to participate in multiple standard-setting procedures and replacement would be possible in case of illness.

During the standard-setting procedure, the tasks for each aspect (cluster) were presented to the experts in a previewer, thus allowing them to view the tasks exactly as the students had seen them. The experts then arrived at an judgement concerning the number of responses that “borderline student” should have answered correctly.[14] For each cluster, the individual experts were asked to indicate how many responses they thought such student should have answered correctly in the cluster if their skills were exactly at the below-level/at-level boundary. The experts could enter their judgements concerning the number of responses that “borderline students” should have answered correctly on a digital standard-setting form (see Figure 11-1). The computer application used during the standard-setting procedure supported the entire standard-setting session (the 3DC application; Koops, 2016; Cito, 2018).

11.5.1 The standard-setting form

The pre-test data were used to create the digital form in which the experts could indicate their judgements concerning the number of correct responses. The pre-test data were used to determine the difficulty of the tasks that were included in the standard-setting procedure. The level of difficulty was determined by placing all of the vmbo tasks along a single ability scale and placing all of the havo/vwo tasks along another ability

[12] Tasks measuring different aspects or those with polytomous responses.
[13] Tasks in which more than half of the concept responses were extremely easy or extremely difficult, or that exhibit very little relationship to the other concept responses.
[14] In the standard 3DC method, each expert was asked to state a cut score instead of the number of correct responses.

scale, using the One Parameter Logistic Model (Verhelst, Glas, & Verstralen, 1995). For each educational stream, simulations for each possible number of correct answers to the full test were used to calculate the expected number of correct answers in a given sub-aspect (cluster). These outcomes were displayed on the digital standard-setting form (see Figure 11-1). In the standard-setting procedure for the DET, the One Parameter Logistic Model was used only for the display on this form. The model was not used for expressing the standard.

An example of the digital standard-setting form is displayed in Figure 11-1, along with fictive judgements (in the circles) by an expert. At the start of a session, the experts were presented with the form without judgements (circles). The correct responses for each sub-aspect are displayed along the horizontal lines of the form. The number of correct responses for a sub-aspect is related to the total number of correct responses. For example, as can be seen on the form in Figure 11-1, a student with a total of about 64 correct responses could be expected to have eight correct responses for "*Conventions*" and five correct responses for "*Spelling*." This form thus showed the experts the relative difficulty of the cluster as compared to the other clusters. The visual relationship between the aspects helped the experts to arrive at realistic judgements.

Figure 11-1. Screenshot of a standard-setting form containing fictive judgements by an expert (visible as the black and red circles). On the left the sub-aspects (clusters) are displayed.

11.5.2 Judgements by the experts

Two rounds of judgements were held for each boundary to be set. The experts started with the below-level/at-level boundary. Based on the tasks, each expert determined the number of responses that a "borderline student" should have answered correctly, and could indicate their judgement on the form (black circles). The first round, in which experts determined their judgements individually, was followed by a group discussion. The group leader used the computer application to provide normative feedback. The group leader displayed the judgements of all of the experts and discussed the similarities and differences between the judgements. The experts were also shown a feedback form that had been developed especially for the DET, as displayed in Figure 11-2. On the form, they could see the range and mode of the judgements. Subsequently, at least two experts were invited to explain their judgements and share their arguments with the group, and a discussion ensued. The discussion served to ensure that the experts had reached their judgements on the same substantive basis. The experts were not necessarily expected to reach the same judgement. After the discussion, the experts had the opportunity to assign a definitive judgement (the red circles in Figure 11-1). After the definitive judgement concerning the below-level/at-level boundary had been provided, the experts started working to reach an judgement of the at-level/above-level boundary in the same manner.

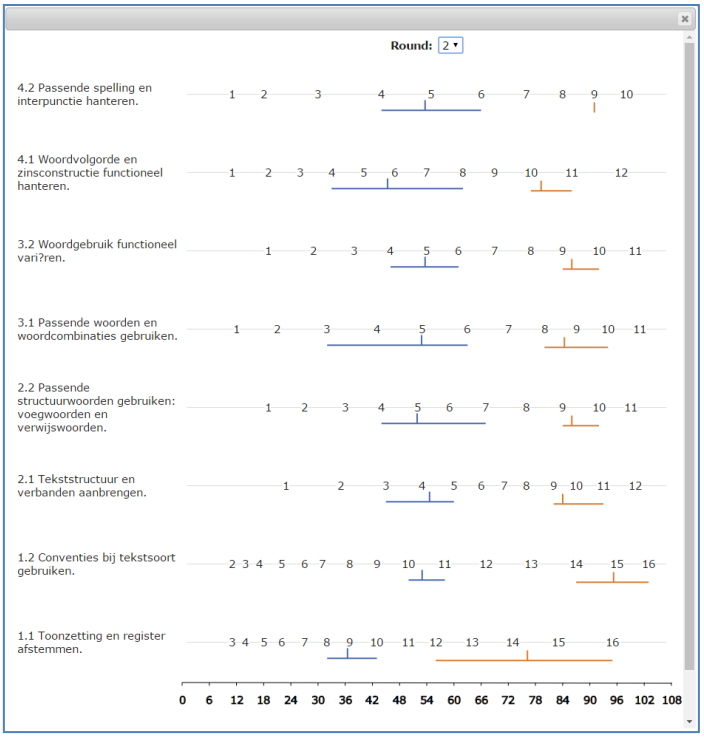


Figure 11-2. Feedback form for the expert judgements, with the judgements of the below-level/at-level boundary in blue and the judgements of the at-level/above-level boundary in red (on the left side the cluster names are displayed, which are the aspect of the student model)

For the DET, the lower boundary is always determined first, followed by the upper boundary. The experience, however, was that the subject-area experts became more lenient throughout the course of the day. In most cases, their first judgements appeared quite strict. A different setup (e.g., in which the first standard was judged again at the end of the day) might be more suitable. In addition, it would be good to examine the utility of setting the upper boundary first, followed by the lower boundary. This might result in a lower standard.

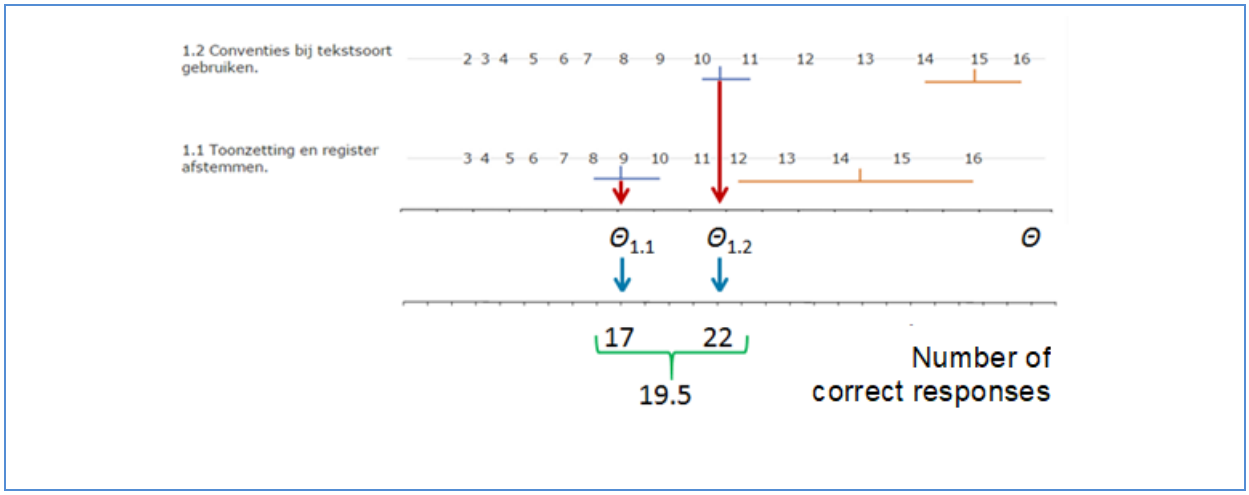
11.6 Confirming the standard

After the standard-setting sessions, the standards were calculated based on the judgements, and a report was written. For the first, limited adaptive administration[15] (2015-2016), the experts' judgements of the aspects (and sub-aspects) were taken as the standard. For the fully adaptive administration (2016-2017), the sub-aspects were used as clusters, which were subsequently used to calculate the standard for the main aspects. To take differences in the relative difficulty of the clusters into account, the judgements were not averaged. Instead, interpolation was used in order to arrive at a standard.

To this end, the experts' average cluster judgements were translated into a ability estimate with linear interpolation ($\theta_{1.1}$ and $\theta_{1.2}$ in Figure 11-3). The ability estimates were subsequently translated into a number of correct responses on a main aspect. In the illustration in Figure 11-3, the ability estimate $\theta_{1.1}$ corresponds to 17 correct responses on the main aspect and $\theta_{1.2}$ with 22 correct responses. The average number of correct responses on the sub-aspects belonging to the main aspect subsequently yielded the standard for the main aspect (in the illustration in Figure 11-3, this average is 19.5).

[15] For the languages, the first adaptive administration yielded only diagnoses on main aspects, and for mathematics, it yielded only diagnoses of the three mathematical sub-aspects in two domains.

Figure 11-3. Illustration of the combination of the judgements into a standard



After each pre-test, a report on the standard-setting procedure was written for the confirmation committees of the Board of Tests and Examinations. The report includes the experts' judgements and the standards, in addition to impact information. The information on impact indicates the number of pre-test students falling into each of the categories (below level, at level, or above level) after the application of the standard. The committee issued a recommendation on the standards. For example, higher or lower standards could be recommended, based on the impact information. After the Board of Tests and Examinations had issued a definitive, formal decision on the standards, they could be applied to the pre-test students and the tasks could be calibrated.

11.7 Conclusion

In two successive years, tasks were pre-tested and standard-setting procedures were conducted (see the pre-test reports: Cito, 2015 and Cito, 2016). Due to the revised assignment on the part of the Ministry of Education, Culture, and Science (see Chapter 1), it was not possible to administer all of the tasks in one large-scale pre-test or to work with only one standard-setting procedure. Further research is needed on the possible impact of the separate standard-setting procedures on the calibration.

The 3DC method was applied successfully for the standard-setting procedures. As is the case with other criterion-related methods, subject-area experts play a central role in the standard-setting procedures with the 3DC method. The experts base their judgements of what could be expected of a borderline student on the tasks, on their experience as teachers, on empirical feedback, and on the substantive description of what a student should be able to do, as described in the interim objectives (SLO, 2012).

The entire process from standard setting to the definitive decision took about four weeks. A norm-related method, as used in the *ad-hoc* calibration (see Chapter 10), would work more quickly, but it is less suitable. With norm-related methods, standards are established according to the *relative* positions of students (Cizek & Bunch, 2007). For example, a norm-related method could involve classifying the 20% lowest-scoring students as being below level. Such norm-related methods are less suitable for indicating how individual students are doing in relation to the objectives that have been set, independent of how other students are doing. In that case, it would be better to use a method based on absolute criteria (e.g., the 3DC method). A criterion-related method is quite appropriate for a diagnostic test aimed at determining the individual learning needs of each student, independent of the performance of other students.

To consider the different structure of the DET (e.g., in contrast to a central examination), the 3DC method was adjusted on a number of points for the DET:

- The clusters were based on student models.
- The IRT model was used only for the standard-setting form, and not to express standards.

- For the DET, the experts were asked to indicate a number of correct responses, and not a cut score.
- The clusters were not combined into a single standard. Instead, two to four cluster judgements were combined into a standard for each aspect.
- Instead of calculating an average score, interpolation was used to calculate the standards.

After several adaptive administrations, it should be possible to evaluate the actual impact of the standards. If such an evaluation were to reveal that substantially more or fewer students than expected were below level, at level, or above level, the standard-setting procedure could be repeated.

11.8 References

Angoff, W.H. (1971). Scale, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Cito (2015). *Diagnostische tussentijdse toets: Verslag pretest 2015 [Diagnostic Educational Test: Pre-test report 2015].* Arnhem: Cito.

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016 [Diagnostic Educational Test: Pre-test report 2016]. Arnhem*: Cito.

Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting*. London: Sage Publications, Inc.

College voor Toetsen en Examens (2014), *Publieksversie Toetswijzer Diagnostische Tussentijdse Toets voor Nederlands, Engels en Wiskunde [Public Version of the Assessment Specification for the Diagnostic Educational Test for Dutch, English and Mathematics].* Utrecht: College voor Toetsen en Examens. Retrieved from https://www.pilotdtt.nl/documenten/publicaties/2014/12/15/toetswijzer-dtt

Feskens, R., Keuning, J., Van Til, A & Verheyen, R. (2014). Prestatiestandaarden voor ERK in het examenjaar: *Een internationaal ijkingsonderzoek [Performance standards for the EFR in the examination year: An international calibration study].* Arnhem: Cito.

Keune, K., & Schouwstra, S. (2014). Standaardbepaling door experts na de eerste pretest van de DTT [Standard setting by experts following the first pre-test of the DET]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [Diagnostic Educational Test: Research 2014]* (pp. 47-52). Arnhem: Cito.

Keuning, J. Straat, J. H., & Feskens, R. C. W. (2017). The Data-Driven Direct Consensus (3DC) procedure: A new approach to standard setting. In S. Blömeke & J. Gustafsson (Eds.), *Standard setting in education: The Nordic countries in an international perspective* (pp. 263-278). Cham, Switzerland: Springer International Publishing.

Koops, J. (2016). *Software and manual 3DC*. http://www.cito.com/our-expertise/implementation/3dc

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Sireci, S.G., Hambleton, R.K., & Pitoniak, M.J. (2004). Setting passing scores on licensure exams using direct consensus. *CLEAR Exam Review*, *15*, 21-25.

SLO: Nationaal expertisecentrum leerplanontwikkeling [Netherlands Institute for Curriculum Development] (April 2012). *Concept-tussendoelen kernvakken onderbouw vo* [Draft interim objectives for core subjects in the lower years of secondary education]. Retrieved from http://www.slo.nl/nieuws/00291/.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: One parameters logistic model. Computer program and manual*. Arnhem, Netherlands: Cito.

12 Psychometric approach within the DET

Daniel van der Palm and Herbert Hoijtink

12.1 Model

The psychometric model for the DET is a latent class (LC) model (Goodman, 1974; Lazarsfeld, 1950), which is combined with “diagnostic hypotheses” (Roelofs & Schouwstra, 2012; Hoijtink & Sies, 2013; Sies, 2014). In the DET, the LC model has a fixed number (three) of latent classes, and it uses prior weights. Use of the latent-class model has increased in recent years with regard to the number of applications in practice, particularly with regard to the number of academic publications. The latent-class model was originally known as “Latent Structure Analysis.” In the field of econometrics, it is also referred to as the “finite-mixture” model.

The LC model is used for situations in which the concepts to be measured are not directly observable. For example, individuals can belong to certain latent classes with regard to personality or political conviction. For example, no direct observation is possible of such characteristics as introversion/extraversion or openness to new experiences. For this reason, an LC model draws upon indicator variables to make an indirect estimate of the latent class to which each individual belongs. This process is also known as “classification.” In the case of the DET, three diagnostic hypotheses (Roelofs & Schouwstra, 2012) are posited with regard to a student’s level of mastery: the student is below level (<), at level (=), or above level (>). The DET thus does not generate a score or a pass/fail result. The goal is to arrive at a diagnosis concerning the extent to which a student has mastered particular aspects (or sub-aspects), such that, if needed, additional time and attention can be devoted to material with which the student is struggling or additional challenge can be offered with regard to material that the student has already mastered.

12.2 Ad-hoc calibration, calibration, and recalibration

Three groups of students can be distinguished: those who are below level, those who are at level, and those who are above level. Calibration is aimed at determining the success probability on items or testlets for each of these three levels. An item is understood as the task or component thereof that generates a response that can be evaluated independently of other responses. In a testlet, a student gives an answer to multiple task components that belong together and that are dependent upon each other. The first calibration that can occur is ad-hoc calibration, which consists of a process that is simpler than the standard calibration. In ad-hoc calibration, a relative standard is applied. One advantage of a relative standard is that it can be applied extremely simply and quickly. The ad-hoc calibration made it possible to provide schools with feedback after each pre-test concerning how their students performed on the pre-test in comparison to other pre-test schools (as described in the pre-test reports: Cito, 2015; Cito, 2016). This calibration was also used in the test and item analyses (see Chapter 10) in order to provide an indication of discriminating power.

In an ad-hoc calibration, for each student an estimate is made per aspect concerning the level group to which the student belongs, based on the percentile in which the student falls in terms of the number of correct responses (sum score; see e.g., Figure 12-1). Students falling within the 20th percentile or lower are classed in the below-level group. Those who are above the 20th percentile but below the 81st percentile are classed in the at-level group, and the rest of the students are classed in the above-level group. Once the groups have been defined, it becomes possible to determine the success probabilities for the items and testlets. This last step of the calibration occurs in the same manner as in a standard calibration. For this reason, we now explain the entire process of standard calibration.

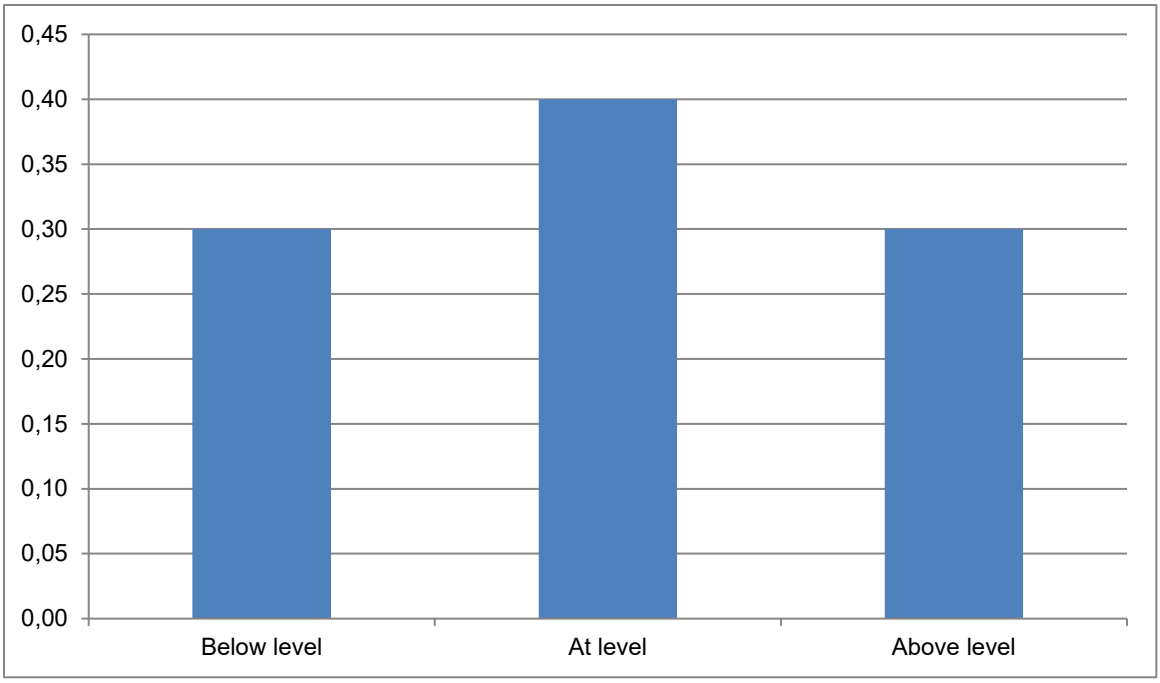


Figure 12-1. Example of a proportional distribution of students in an ad-hoc calibration based on the number of correct answers

An absolute criterion-related method is used in a standard calibration. The standards for each subject are determined, based on consultations with experts and according to the data-driven direct consensus method (see Chapter 11). There are two standards for each aspect: the number of correct responses that the experts regard as being consistent with (a) students whose mastery of the subject is just barely at level and (b) students whose mastery is just above level. After the standards have been determined, a determination is made for each student concerning whether the student is below, at, or above level. The essential difference between ad-hoc calibration and standard calibration thus concerns whether the boundary values are based on relative percentiles or on absolute standards. For relative percentiles (ad-hoc calibration), the standard depends upon the performance of other students. Absolute standards that are used along with subject-area experts are based on the interim objectives, test content, and empirical feedback (see Chapter 11).

After determining the three groups of students, the parameters of the LC model are estimated. There are three sets of parameters: (1) the success probabilities of separate items, (2) the success probabilities of testlets, and (3) the prior-model probabilities,. The success probabilities of separate items can simply be observed within the three groups of students (below level, at level, and above level). For example, the success probability for item *j* within the above-level group is calculated as follows,

$$P_j(>) = \frac{N_j(>)}{N_j(>) + M_j(>)}$$

with $N_j(>)$ standing for the number of students in the above-level group who have answered item j correctly and $M_j(>)$ standing for the number of above-level students who have answered item *j* incorrectly.

In a testlet, a student gives an answer to a set of tasks or components thereof that belong together and that are dependent upon each other. For this reason, it is important to consider the fact that the various answers to this set of tasks could display internal statistical dependence. The solution to this problem is to model the success probabilities on testlets at the level of response patterns. A testlet that generates three responses has 2^3 = 8 possible response patterns, and the probability that a student will give each of the eight response patterns must be estimated for each of the three level groups (below-level, at-level, above-level). The number of possible response patterns increases rapidly along with the number of responses in a

testlet, and far from every possible response pattern will be observed in a given administration. For this reason, a second-order log-linear model is used for testlets, thus making it possible to estimate the probability of each possible response pattern. Without this information, it would be impossible to include the entire series of possible response patterns in the simulation procedure (described in paragraph 12.3.1). When more than four responses actually belong to a single testlet, these responses are not interpreted as a testlet, but processed as separate items.

Finally, in addition to the two sets of success probabilities, the prior-model probabilities are needed in order to estimate the level groups of individual students based on their response patterns. These probabilities consist of three weights that make it possible to adjust tendencies in the diagnoses. For example, one tendency could be that fewer below-level students are diagnosed correctly than is the case for at-level students. The procedure starts by using “non-informative priors,” such that each of the three weights equals 1/3, thus not yet requiring any adjustment. Further details concerning prior-model probabilities are provided in Section 12.4.

12.2.1 Recalibration

The result of a calibration is a collection of estimated success probabilities for the items and the estimated probablities of each response pattern for the testlets. In principle, when new administration data from previously calibrated items and testlets become available, the calibration should be performed again based on all available data in order to reduce the influence of sampling fluctuations (standard errors of model parameters decrease as sample size increases). At this point, it is important to investigate whether the results of a calibration based on the new data (parameters of items and testlets) differ excessively from the previously estimated parameters. Because this process involves samples, slight differences could be expected between the estimated probabilities from different test administrations, but systematic shifts (also referred to as “parameter drift”) could occur as well (see also, Maas, 2017). If a shift occurs, the manner in which this discrepancy is addressed should be examined in greater detail. If the results are reasonably similar, a recalibration can be performed based on all available data. The output from this recalibration is then used in the composition of new tests.

12.2.2 Imputation

Due to the adaptive structure of the DET, not all of the tasks are presented to every student, and this results in missing values. For example, within the group of below-level students, certain tasks will be answered rarely, if ever, as they appear later in the test and are offered only to students who probably belon to the at-level group. Another possible cause of missing values is that not all of the tasks are presented to some students, as their diagnoses were already certain enough (see Chapter 13). If the missing values are not addressed, there would be a few tasks for which very few, if any responses would be observed within one or more of the three level groups. In other words, the estimated success probability of this level group cannot be estimated or is highly unstable, as it is based on a small number of observations.

Imputation can make it possible to estimate responses for students with missing values, thereby enhancing the stability of the estimated probabilities. In the case of the DET, an LC model is used to estimate the joint probability distribution of all responses. Such an estimated probability distribution can be used to generate responses for students with missing values.

The number of latent classes used for the imputation is determined through iteration based on an information criterion. An information criterion makes it possible to select a model, with the model fit being weighed against the number of parameters. For each dataset, a LC model with two classes is estimated first. The number of classes is increased in each iteration until the improvement in model fit no longer justifies the increase in the number of parameters.

If the LC model is estimated for imputation, calculations are to determine the probability that each student with at least one missing value would belong to each of the three level groups, and students are randomly assigned to categories based on these probabilities. The conditional response probabilities are then used to generate responses to items with missing values for each student.

12.3 Optimization of the block arrangement using simulation

With three sets of parameters, the LC model is completely defined, in principle. Nevertheless, the exact tasks that will be presented to a given student are not determined. In theory, it would be possible to present all tasks to each student, regardless of that student's level of mastery. In practice, however, this would be an extremely time-consuming process, and it would impose an undue burden on the student. For this reason, an adaptive design was used (see also Chapter 13). An adaptive design for the DET consists of 2, 3, or 4 layers[16] of tasks, with each layer containing a separate block of tasks for students who are probably below level, students who are probably at level, and students who are probably above level.

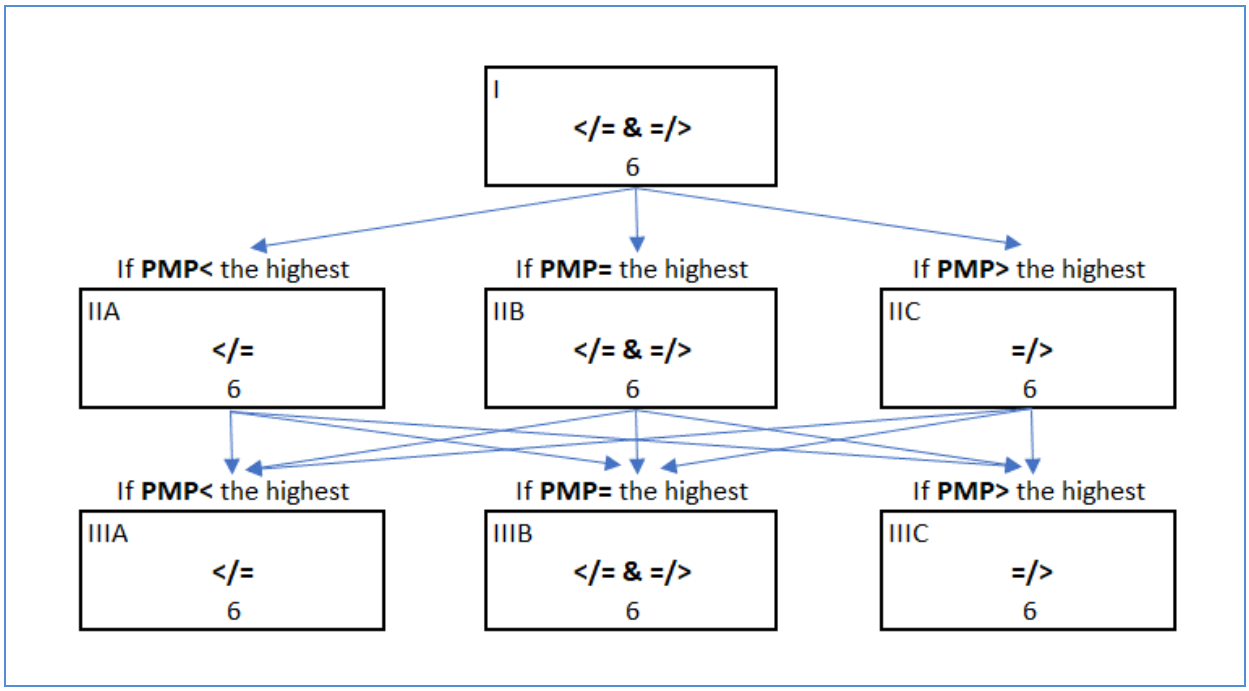


Figure 12-2. The adaptive procedure with blocks (Hoijtink & Sies, 2014)

As demonstrated in Figure 12-2, all students start by completing the same block of tasks for a given aspect. This first set of tasks should thus make a good distinction between students who are below and at level, as well as between those who are at and above level. After proceeding through the first block, an estimate is made of the level-group the student belongs to. This is calculated as follows:

> The prior-model probabilities and success probabilities of each of the three level groups are estimated, thus making it possible to calculate the probability that a given response pattern will be observed in each level group (Hoijtink, Béland, & Vermeulen, 2012): the prior-model probability for a level group multiplied by the probability that a given response for the first item will be observed and so forth, thus generating the sum product across all items that have been answered by a student. If each of the three probabilities are divided by the sum of these three probabilities, this produces the "posterior membership probabilities" (PMP). These PMP's indicate the probability that a student belongs to a specific level group, given the student's response pattern.

After the first block, each student is tentatively assigned to the level group for which the student has the greatest PMP (i.e., the most probable level group). In the second layer, the student is subsequently given a block of tasks belonging to this level group. For example, students who are probably below level will be presented with tasks that are good at distinguishing between students who are below and at level (left block in Figure 12-2). Students who are probably above level will be presented with tasks that are good at distinguishing between at level and above level (right block in Figure 12-2). This process is repeated at the

[16] In 2016, there were two layers and, in 2017, there were three layers without further testing (mathematics) or with further testing (languages).

end of the second layer (and at the end of the third layer, if there are four layers). At the end of the test, the PMP values (i.e., the assessment outcome) are reported. At the end of the test, we know the probability that a student belongs to the below-level group (PMP<), based on that student's response pattern, along with the probability that the student belongs to the at-level group (PMP=) and to the above-level group (PMP>). Each student is placed in the most probable level group (i.e., the one with the highest PMP). For example, a student with the highest probability of belonging to the at-level group will receive a diagnosis of "at level."

In technical terms, this completes the test administration for one main aspect. At that point, however, the manner in which the tasks should be distributed across the various blocks is not yet clear. A simulation procedure is used to answer this question.

12.3.1 Simulation

Simulation makes it possible to search for the best possible distribution of the tasks and responses across the blocks. First, it is necessary to define the criteria to be adopted in order to assess the quality of a given arrangement. Two criteria were taken into consideration: (1) the size of the probabilities of correct diagnoses, given the true level of each student, and (2) the balance among the three probabilities of a correct diagnosis.

The nine possible outcomes of the test are displayed in Table 12-1, distinguishing between a student's true level group and the level group estimated at the end of the test administration. In the hypothetical example, 88% of the below-level students receive the proper diagnosis of below level, with the rest receiving incorrect diagnoses: 9% of the incorrect diagnoses are for at level, and 3% are for above level. Of the at-level students, 87% are diagnosed correctly, as are 91% of the above-level students.

Table 12-1. Example of a classification table

	Estimated level		
True level	Below	At	Above
Below	0.88	0.09	0.03
At	0.07	0.87	0.06
Above	0.01	0.08	0.91

The first criterion – the probability of a correct diagnosis – involves attempting to maximize the sum of the diagonal of this 3x3 table, given that these are the outcomes in which the true and estimate level groups converge. The simulation was developed to meet this first criterium and the second (the balance among the three probabilities of a correct diagnosis) as closely as possible, and it works as follows.

1) The tasks are randomly distributed across the blocks.
    a. For example, in the case of three layers (without further testing), there would be seven blocks, as displayed in Figure 12-2. If there must be seven blocks, the tasks are distributed across five blocks, as the middle at-level block in the second and third layers are formed of tasks from the "below-level" and "above-level" blocks. To this end, for each sub-aspect half of the tasks are randomly taken from the "below-level" block, with the other half being taken from the "above-level" block.
    b. The number of tasks that can be assigned to a block is restricted according to the available testing time. The maximum amount of testing time available for each main aspect is known at the start. For example, if three hours are available for the entire administration and there are four main aspects and seed tasks (see paragraph 13.5), 2160 seconds will be available for each main aspect (10,800 seconds/5). The amount of testing time available for each main aspect is divided by the number of blocks that a student must complete (layers) in order to arrive at the amount of time available for each block. The time available for each block (a student has complete) is then divided by the amount of response time available for each task in order to arrive at a maximum number of tasks for each block. For example, for a testing time for each main aspect of 2100 seconds and a three-layer design with further testing, a student would receive four blocks: 2100/(4 blocks) = 525 seconds per block, 525/(50 seconds response time needed for each task) = 10

tasks. A student will thus be presented with a maximum of 10 tasks in each block. This could mean that not all of the available tasks will actually appear in the test. The response time needed for each task was calculated on the basis of the observed response times during the previous administration.

2) For the block arrangement, 10,000 simulated below-level students are created, along with 10,000 at-level students and 10,000 above-level students. These simulated students proceed through the test with the block arrangement that has been created. Answers are generated for the tasks that the simulated students receive according to the block arrangement, based on the success probabilities, conditional upon the level group. The level group of each simulated student is estimated at the end of each layer. At the end of the test, the diagnosis is known, and it becomes obvious which of the nine possible outcome groups the simulated student has been classed in (see Table 12-1 in the classification table). The frequency of the relevant outcome in the 3x3 table is moved up by one.
3) The creation and simulation of this block-arrangement procedure continues until 100 block arrangements that meet the requirements have been found. There are three restrictions:
    a. The time restriction
    b. The requirement that each block must have at least one response for each sub-aspect
    c. The requirement that the number of responses are distributed evenly enough across the blocks
4) The sum of the three correct-diagnosis likelihoods is calculated for each of these 100 arrangements, and the 10 arrangements with the highest sum of the diagonal are retained.
5) Finally, the prior-model probabilities are optimized, thereby minimizing differences in the probabilities of correct diagnoses (see Section 12.4)

These 10 best solutions are fed back to the assessment experts, who consult with the psychometrics department to identify the solution that should be preferred (see Chapter 13).

Task groups in the block arrangement

In some cases, the optimizations must take into account the presence of task groups. These groups are composed of tasks that are presented on different screens and that belong together (e.g., three tasks concerning the same text). This means that these tasks must be presented at the same time in a specific order. The search algorithm for finding an optimal distribution takes task groups into consideration by bundling such tasks in advance within a temporary structure. If there are two tasks, each with one response, and if these tasks belong to the same task group, the simulation temporarily makes them a single task with two responses. This is an issue only in the part of the simulation examining whether random block arrangements meet the various restrictions. Once a block arrangement that meets the restrictions has been found, these bundles are returned to their original form, and there is at least a guarantee that the relevant tasks will be placed in direct succession in the same block.

12.3.2 Exit probabilities

The adaptive design also takes into consideration the fact that a student need not complete all of the blocks if the diagnosis is already quite certain (e.g., if the student has completed the tasks consistently correctly or incorrectly; see Chapter 13). In essence, the student may exit the test. An exit probability is a criterion that can be used at the end of each layer to determine whether a student can exit because sufficient information has already been collected about that student's level. This exit probability can be calculated separately for each layer. In the technical structure of the adaptive module, however, the choice was made to have only one set of exit probabilities for each main aspect (and one set for each sub-aspect). The set of exit probabilities is calculated according to the tasks in the first layer, as they yield the most conservative values. These exit probabilities are subsequently applied after each layer.

The determination of exit probabilities proceeds from the assumption that items and testlets in a diagnostic educational test have discriminating power. The discriminating power varies from item to item and from testlet to testlet. Nevertheless, each additional task that is presented to the student increases the certainty of the estimated level (unless the success probabilities are equal for the various levels, but this can be detected during the calibration, and it applies to only a very limited number of items).

For the DET model, the larger the number of test items that a student completes, the more likely it will be that a correct diagnosis will be made (i.e., the estimated level for a student is equal to the true level of that student). Theoretically, the probability of a correct diagnosis approaches 1 as the number of test items approaches infinity, given that the influence of chance is increasingly eliminated under these circumstances. In practice, the number of items is obviously limited, and the influence of chance is greater. We will thus not easily arrive at an assessment that can make a perfect distinction.

Even if the assessment does not make perfect distinctions, however, the observation of a slightly different response pattern (given the level of a specific student) is more likely than is the observation of a widely differing response pattern. In other words, the probability that a student in the below-level group will complete a very difficult task correctly by chance (e.g., by guessing or due to a particularly insightful moment) is greater than the probability that the same student would be able to answer 10 very difficult tasks correctly. This also means that, in general, if a faulty diagnosis is made, the difference between the probability (PMP) of the estimated level group and the probability (PMP) of the true level group is usually relatively small. For example, if a below-level student receives a faulty at-level diagnosis, the assessment outcome for that student is more likely to be [$PMP_{<}$ = 0.3; $PMP_{=}$ = 0.5; $PMP_{>}$ = 0.2] than it is to be [$PMP_{<}$ = 0.1; $PMP_{=}$ = 0.2; $PMP_{>}$ = 0.7]. It can also be concluded that, on average, a higher proportion of incorrect diagnoses will occur in the collection of PMP vectors that do not differ much from a uniform distribution [1/3, 1/3, 1/3] and that fewer errors will be found among students with PMP vectors having a strong tendency toward a specific level (e.g. [0.7, 0.2, 0.1]).

A rule was implemented to prevent early exit unless the percentage of correct diagnoses in the group of students exiting is 95% or higher (see Chapter 13), in addition to preventing exit until after the full first layer has been completed. To identify an optimal block arrangement in the simulation procedure, a calculation was performed to determine when this 95% requirement could be met for each of the three level groups. In any case, the PMP of the simulated students (e.g., those estimated as being below level) will be greater than 1/3 and can increase to nearly 1.0. As noted previously, more faulty diagnoses could be expected to occur among students whose PMP values for “below level” are the greatest, but only slightly. For this reason, students who are estimated as being “below level” are ranked from low to high according to their PMP values (<). In principle, the number of incorrect diagnoses could already be lower than 5% (e.g., if the first layer contains tasks that are very good at distinguishing between below-level and at-level students and between at-level and above-level students). In such cases, the exit probability is set at.01, such that nearly everyone would exit after the first layer if they were to be diagnosed as being below level. It is also possible that the 95% requirement cannot be met at all, due tasks of lesser quality. In such a case, the exit probability would be set at.99. For the cases between these two extremes, the minimum PMP in the simulation procedure is continually increased until the group with PMP values above a specific PMP value consists of at least 95% correctly diagnosed students. If the 95% requirement is met, the exit probability is set equal to this PMP value.

12.4 Optimization of the prior-model probabilities

As briefly stated before, prior-model probabilities are the weights that can be used to adjust tendencies in the probabilities of correct diagnoses. In the DET there are three levels, and thus three prior weights. Once a certain block arrangement has been found in the simulation procedure, we can be certain that, in any case, the probabilities of correct diagnoses was highest for this arrangement (within the set of block arrangements examined). Nevertheless, this does not guarantee that the probability of correct diagnoses is the same for each of the three levels. The differences between these probabilities are usually not particularly large, as there are usually many alternative arrangements that do not display this derogation and, in the search algorithm, these alternative arrangements might displace any arrangements that display flaws.

Because we are working with discrete units (items and testlets) that must be divided, however, the three likelihoods of correct diagnoses will never be exactly equal. Prior-model probabilities are used in order to realize this as closely as possible. For the 10 best solutions, the following steps are used to optimize the prior-model probabilities and to make the likelihood of correct classification as equal as possible.

(1) Take the probabilities of correct classification for a given arrangement, $pr_<, pr_= \text{ en } pr_>$, and calculate the vector **m** as follows: $\mathbf{m} = \left(pr_>, \frac{pr_>}{pr_=/pr_<}, pr_<\right)$
(2) Normalize vector **m** by dividing each of the three elements by the sum of **m**
(3) Multiply the normalized vector **m** by the old prior-model probabilities, and normalize the result again, in order to arrive at a new set of prior-model probabilities
(4) Recalculate $pr_<, pr_= \text{ en } pr_>$, given the new prior-model probabilities, and return to Step 1 as long as the prior-model probabilities in an iteration are greater than the lower boundary that has been set.

12.5 References

Cito (2015). *Diagnostische tussentijdse toets: Verslag pretest 2015 [Diagnostic Educational Test: Pre-test report 2015]. Arnhem*: Cito.

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016 [Diagnostic Educational Test: Pre-test report 2016].* Arnhem: Cito.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*, 215-231.

Hoijtink, H., & Sies, A. (2013). De psychometrie van de diagnostisch tussentijdse toets [The psychometrics of the DET]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2013 [The Diagnostic Educational Test: Research 2013]* (pp. 13-62). Arnhem: Cito.

Hoijtink, H., & Sies, A. (2014). Adaptieve procedure met itemblokken [Adaptive procedure with item blocks]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [The Diagnostic Educational Test: Research 2014] (*pp. 33-40). Arnhem: Cito.

Hoijtink, H., Béland, S. en Vermeulen, J. A. (2014). Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychological Methods*, *19*(1), 21-38.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction*, pp. 361-412. Princeton: Princeton University Press.

Maas, L. (2017). *Potential impact of item parameter drift on diagnostic accuracy in educational testing* [Master's Thesis]. Utrecht: Utrecht University.

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

Sies, A. (2014). *Cognitive Diagnostic Testing Using Calibrated Hypotheses* [Master's Thesis]. Utrecht: Utrecht University.

13 Adaptivity in the DET

Sanneke Schouwstra, Peter van Os and Joke Hofstee

The DET serves a formative function. The DET should therefore provide more detailed information than a general ability score, thus making it possible to tailor the learning process to individual learning needs. For this reason, the DET is based on student models that describe all of the knowledge and skills aspects that are needed to perform at a given level (see Chapter 1). The DET is intended to diagnose all such knowledge and skills aspects (see Chapters 4, 5, and 7). For the first skills operationalized (Dutch writing skills, English writing skills and mathematics), it should be possible to diagnose 8-13 sub-aspects. Such a large number of diagnoses cannot be generated unless the assessment is adaptive. Computerized adaptive testing (CAT) can increase the effectiveness of testing by determining after each task how much we already know and which tasks must still be presented in order to arrive at an accurate diagnosis.

The form of adaptivity is tailored to the function of the test. Adaptivity can be used for a variety of objectives. In educational testing, adaptivity is often used to increase the accuracy of measurements across the entire ability distribution or to allow accurate measurement of the ability distribution, even at the extremes. In psychological measurement instruments, adaptivity is used primarily to reduce the number of questions, and thus the testing time (see e.g., Hol, 2006). For the DET, the primary reason for using adaptivity is to make it possible to arrive at more accurate diagnoses in the same amount of time needed for a linear administration. In the DET, we would like to be able to estimate whether the student is below, at, or above level for each aspect of a student model (see Roelofs & Schouwstra, 2012; Sies, 2014). This process does not use models based on item-response theory, as is customary in computerized adaptive testing (Eggen, 2012), but the model described in Chapter 12.

13.1 Grades of adaptivity

Various grades of adaptivity are combined in the DET (see Figure 13-1). First, the DET was developed for five different educational streams (the second year of vmbo-bb, vmbo-kb and vmbo-gt and the third year of havo and vwo), with a separate version for each educational stream. Overlap in tasks was created between the versions of adjacent educational streams, as the original assignment called also for enabling an indication about moving up to the next educational stream (see Roelofs & Schouwstra, 2012). This requirement was eliminated in the revised assignment from the Ministry of Education, Culture, and Science, in part because it would require pre-testing in the second year of havo (in which the DET is not administered; see Chapter 1).

Second, in addition to the two different versions, there are two stages in the measurement. Analyses after the preliminary study and the first pre-test indicated that many hours of testing would be needed in order to diagnose a complete student model. For this reason, it was decided to work in two phases. In the first phase, all of the main aspects are diagnosed. In the second stage, the further-testing stage, sub-aspects are diagnosed only if necessary and only if there is testing time remaining (Hoijtink & Sies, 2013). Section 13.4.2 includes a discussion of the algorithm for further testing, which is used to determine the order in which the sub-aspects will be diagnosed further.

Finally, the most refined adaptivity occurs in the diagnosis of the main aspects. In order to arrive at the desired diagnoses more quickly, an adaptive procedure involving task blocks is used in the first phase.

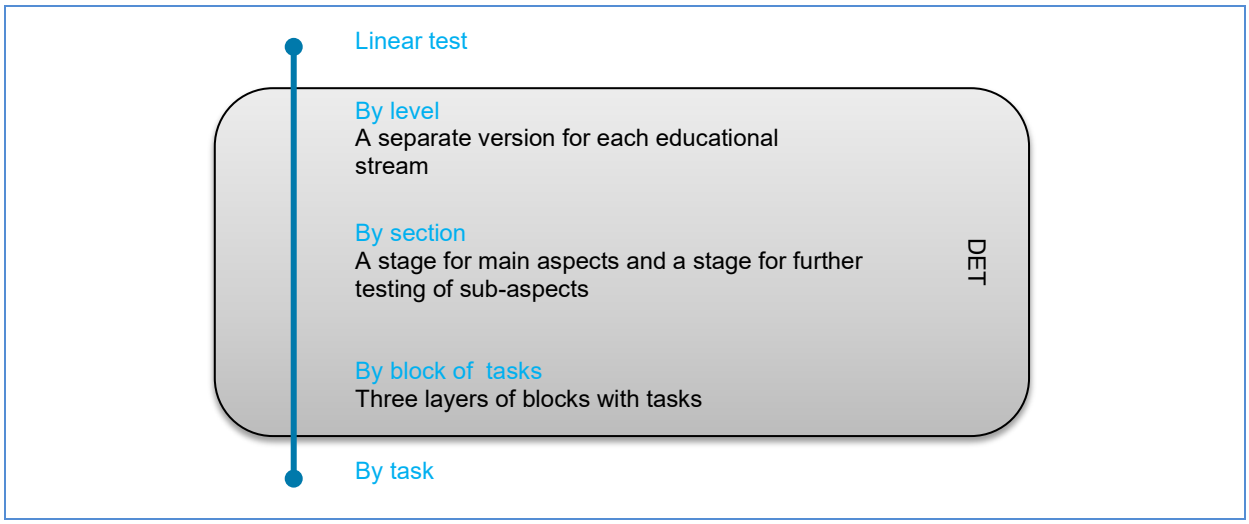


Figure 13-1. Grades of adaptivity in the DET

13.2 Adaptive procedure with blocks

For reading comprehension, as well as for Dutch writing skills, multiple tasks are sometimes included for a single text. These tasks obviously belong and should be presented together. The DET thus contains existing blocks of tasks (task groups). This is nevertheless not the most important reason for working with blocks of tasks. Recalibration of the parameters of tasks would not be possible without a new pre-test (Zwitser & Maris, 2015), even though it will always be necessary in the course of time. The use of a block structure does make it possible to perform recalibration based on adaptive administration.

Recalibration is needed when the administration group responds to tasks differently than they had done on the pre-test (or previous administrations), because of which the parameters of the task change ("item-parameter drift;" see also Glas, 2000). After a few years, the administration group will respond differently because the students have practiced before the measurement, because the teaching has changed, or because the tasks have become familiar (Maas, 2017). Different reactions can be expected even earlier if the manner of administration has been changed, if the pre-test group was very small, or if it was not representative of the student population. Recalibration was needed after the revised assignment from the Ministry of Education, Culture, and Science (see Chapter 1) which involved working with a group of schools participating voluntarily, due to the smaller pre-test group (500-1000 observations for each task).

The use of task blocks also compensates for other disadvantages of task-level adaptivity (see e.g., Hendrickson, 2007). Experts are better able to assess the quality of assessments if blocks are used. Assessment experts can inspect the content of the blocks in order to prevent context effects, to balance the content of the tasks, and to determine the proper order. The extent to which the tasks are seen by students (i.e., "item exposure") is also easy to inspect, and adaptivity based on blocks is technically less demanding.

Simulation studies have indicated that the adaptive procedure with blocks works well when the distribution of the success probability is used as a guideline instead of the difficulty of the tasks. Very little information is lost relative to the adaptive procedure at the task level (Hoijtink & Sies, 2014).

13.2.1 Block design

The block design is drawn in Figure 13-2. The block in the first layer (A) is presented to all students. Thereafter, one block from the second layer is presented:

- Block B is presented to the students most likely to be "below level" based on the first layer (Block A),
- Block C is presented to the students most likely to be "at level" based on the first layer (Block A) and
- Block D is presented to the students most likely to be "above level" based on the first layer (Block A).

After the second block, one block from the final (third) layer is presented:

- Block E is presented to the students most likely to be “below level” based on the second layer,
- Block F is presented to the students most likely to be “at level” based on the second layer, and
- Block G is presented to the students most likely to be “above level” based on the second layer.

After proceeding through the third layer, the student begins the second stage of further testing. A further-testing block containing tasks is provided for each sub-aspect.

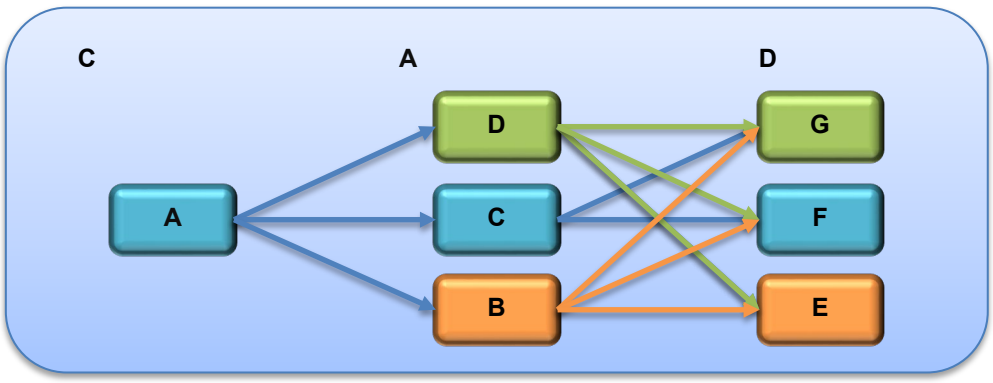


Figure 13-2 Block design

In the first, limited, adaptive administration of 2016, only the first two layers were created for each main aspect, and no further testing was conducted (Cito, 2015). The last administration of the pilot study in 2017 included the implementation of the entire three-layer adaptive design and further testing (Cito, 2016).

13.2.2 Creating blocks

The blocks were initially composed automatically. The tasks belonging to a main aspect were randomly distributed across the blocks 1000 times (see Chapter 12; Cito, 2016). Simulated data were then used to assess the accuracy of the diagnoses for each of these 1000 solutions. The 10 solutions with the most accurate diagnoses at the main-aspect level were submitted to psychometricians. The pyschometricians examined a total of 600 solutions (3 subjects x 4 main aspects x 5 educational streams x 10 solutions) for an adaptive administration of Dutch writing skills, English writing skills, and mathematics. In addition to the accuracy of the diagnosis of the main aspect, attention was devoted to the accuracy of the sub-aspect diagnoses, noting the distribution of responses belonging to the various aspects across the blocks.

The assessment experts then examined the advised solutions (20 for each subject). In this process, the assessment experts were alert to potential context effects, order effects, and imbalanced content. If desired, the arrangement of blocks could be adjusted by hand, obviously using simulations to inspect the accuracy. For example, for English writing skills in the vmbo-bb educational stream, many tasks assessing the formulation of a heading in a letter appeared in a single block after the automatically composed block arrangement (for the 2017 administration), even though several different types of tasks were used. The solution was then adjusted by hand. These heading tasks were distributed equally by hand across the blocks of Main Aspect 1, *Tuning to audience and objective*, and they were separated by considerable distance from each other within the blocks.

13.3 Adaptive course

In the first limited adaptive administration of 2016, students completed the tasks by main aspect. They started with the blocks for the first main aspect. They then completed the blocks for the second main

aspect, followed by the blocks for the third and, finally, the fourth main aspect (Cito, 2016). The confirmation committee determined that, in the 2017 adaptive prototype, students would proceed through the measurement by layer, as this would make the measurement more varied for students (see Figure 13-3). In this variant, all of the students started by completing the first block for each main aspect, followed by several seeding tasks (see Section 13.5). They then completed a second block for each main aspect and, finally, a third block (see Figure 13-3). After completing the entire first stage, they started on the further-testing blocks. The order in which this was done was determined by the algorithm for further testing (see Section 13.4.2).

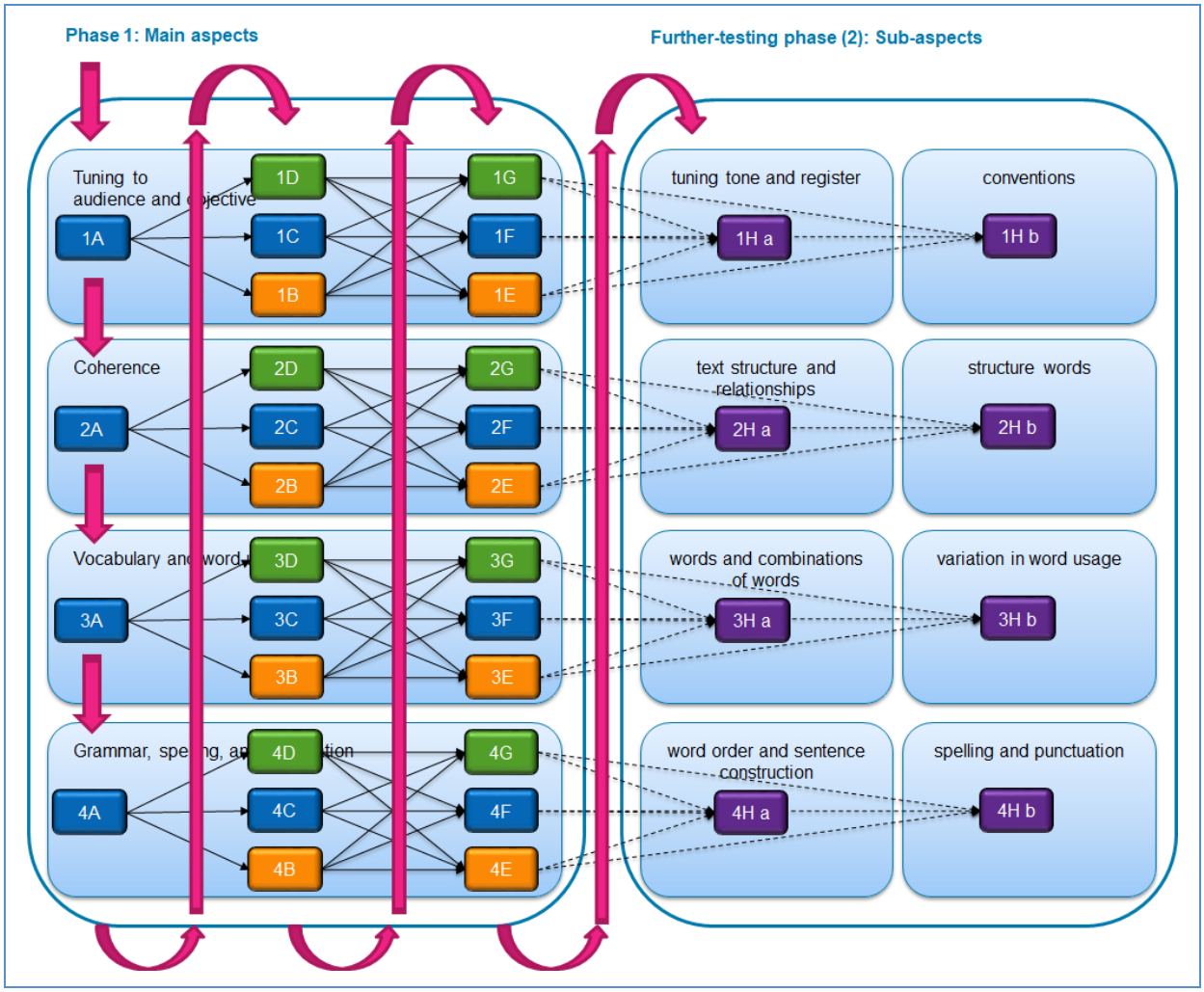


Figure 13-3. Order of the adaptive test (pink arrows) for English writing skills in 2017

Two disadvantages were associated with proceeding by layer (as done in 2017). First, students spent much more time working all on exactly the same tasks. The first block was completed by all students. In the layer structure, all students start with the same block for each main aspect. Only when they have completed the first four blocks of the first layer will the first adaptive decision take place (Block B, C, or D of the first main aspect, see Figure 13-3). It thus makes it slightly easier to cheat.

A second disadvantage is that it takes longer before the initial diagnosis can be made. In a course that proceeds by main aspect, the initial diagnosis (for the first aspect) can be made after three blocks. In a course that proceeds by layer, the initial diagnosis cannot take place until after nine blocks. That creates a risk if the entire time has not been used, as students will not get a complete report in that case.

13.4 Adaptive rules

Three adaptive rules were implemented: An exit rule, a rule for further-testing order, and a rule for the usability of the outcomes.

13.4.1 Rule for exiting or skipping blocks

It is possible for the diagnosis to be quite certain before the third layer is reached (e.g., if the student has completed the tasks consistently correctly or incorrectly). Once the diagnoses for the main aspects and their associated sub-aspects are sufficiently certain, the student will not receive any more tasks on the main and sub-aspects. In the first adaptive administration (2017), a fixed criterion of 0.95 was used. It would be better, however, to have the criterion depend upon the desired percentage of correct diagnoses. For this reason, a procedure was developed for determining the probability at which the desired level of accuracy (percentage of correct diagnoses) has been reached.

Simulations were used to define boundary values for each aspect. For each aspect, a calculation was performed to determine the probability (posterior model probability, see Chapter 12) at which 95% of the students received correct diagnoses after the first block of tasks. After each answer, the probabilities (posterior model probabilities) for all diagnoses were adjusted. If the probabilities for the main aspect and the associated sub-aspects exceeded the boundary values, the student was not presented with any more blocks of tasks for these main and sub-aspects.

The probability at which 95% of students were diagnosed correctly (the boundary value) was usually less than.95. We observed that, for many aspects, 95% of all below-level students and 95% of all above-level students were diagnosed correctly at a probability much less than 0.90. For the at-level students, the boundary value at which 95% of all at-level students were diagnosed correctly was usually between.90 and.95 (see Appendix 17.2 for all boundary values).

13.4.2 Rule for further-testing order

After all of the main aspects have been diagnosed in the first stage, the further-testing stage begins. If the planned testing time (3 hours) has not yet elapsed and if the administration has not yet been stopped by the administration coordinator, it is possible to continue to the assessment of sub-aspects. As described previously, a sub-aspect does not need to be assessed further if the diagnosis is already sufficiently certain. This is the case if the probability of the diagnoses for the sub-aspect already exceed the boundary value, at which 95% of all students are diagnosed correctly. All of the other sub-aspects are eligible for further testing.

Proceeding from the notion that all students, including the excellent students, can maximize their learning outcomes by addressing relative points for improvement, an algorithm emphasizing relative points for improvement was developed (Hoijtink & Schouwstra, 2014). An indication occurs if there is a substantial chance that the “student’s mastery of the sub-aspect is in need of relative improvement.” To this end, the estimated posterior model probabilities after the first stage (i.e., after the student has completed or been allowed to skip three blocks of tasks for each main aspect) are examined.

- If it is estimated that most of the sub-aspects will be above level, the sub-aspect that is least likely to be above level (i.e., the one with the lowest PMP>) will be subjected to further testing first.
- If it is estimated that most of the sub-aspects will be at level, the sub-aspect that is most likely to be below level (i.e., the one with the greatest PMP<) will be subjected to further testing first.
- If it is estimated that most of the sub-aspects will be below level, the sub-aspect that is least likely to be below level (i.e., the one with the lowest PMP<) will be subjected to further testing first.

The second stage ends when the available administration time has elapsed or the diagnoses for all sub-aspects are sufficiently certain. Alternative algorithms are obviously conceivable. One example could be an algorithm that considers strengths as well as the relative points for improvement. In that case, an indication for further testing would also occur if the student’s mastery of the sub-aspect were to be relatively strong.

13.4.3 Rule for the usability of outcomes

In the adaptive module, the probabilities (posterior model probabilities) of all diagnoses for all aspects and sub-aspects are adjusted. The adaptive module renders an estimate for a particular aspect after one response. To prevent the reporting of diagnoses based on too few responses, a rule was formulated for the usability of the estimates.

The evaluation of the block arrangements includes the assessment of the accuracy of the diagnoses (see Section 13.2.2). The current guideline is that, if less than 60% of the students who are below, at, or above level are diagnosed correctly, the outcomes cannot be used for reporting on any student. If this is the case, a statement may be entered into the driver file for the adaptive test that the outcomes of the aspect in question are not useable within that block arrangement.

For all aspects for which the simulations indicate that more than 60% of the students have been diagnosed correctly, the following three rules are implemented in the adaptive module:

- If the probability of the diagnosis exceeds the boundary value, the assessment outcome is useful, and it can be reported.
- If the probability of the diagnosis is below the boundary value and if the student has proceeded through all three layers of the first stage, the diagnosis of the particular main aspect is useable, and it can be reported.
- If the probability of the diagnosis is below the boundary value and if the student has completed the entire further-testing block, the diagnosis of the particular sub-aspect is useable, and it can be reported.

Therefore, if the student terminates the administration very early, it is possible that the student will not have provided enough responses for any aspect. In that case, the student would not receive any report.

13.5 Seeding

The adaptive design also considers the possibility of seeding new tasks. The inclusion of new tasks in the adaptive test makes it possible to use the adaptive test to determine the properties of these tasks, as well as to calibrate the tasks, if they have been approved (see Chapter 12). The advantage of seeding is that it eliminates the necessity of a large-scale pre-test in order to calibrate new tasks. Each year, seeding results in the addition of new tasks, which can be used to refresh and maintain the item bank.

To ensure that all students, regardless of level, completed the seed tasks, the seed tasks are presented after the first block. A notation must be made in a file in advance (see the following section) indicating how many seed tasks a student will receive for each sub-aspect. The number of seed tasks that can be included usually depends upon the available testing time and the expected number of observations. Seeding should not demand too much administration time, and at least 500 observations are needed in order to calibrate the tasks after administration. In the adaptive administration of 2017, all of the students received the same seed tasks, although it would also be possible for each student to receive only a randomly selected number of tasks.

13.6 Adaptive architecture

An adaptive model has been programmed for the DET. This module is included along with a test. A “driver file” is also included, containing all of the settings for the adaptive administration. Separate tools have been programmed for creating and testing the driver files (see Figure 13-4) and for visualizing the adaptive design. The module and tools come with documentation, so that other parties can also use them.

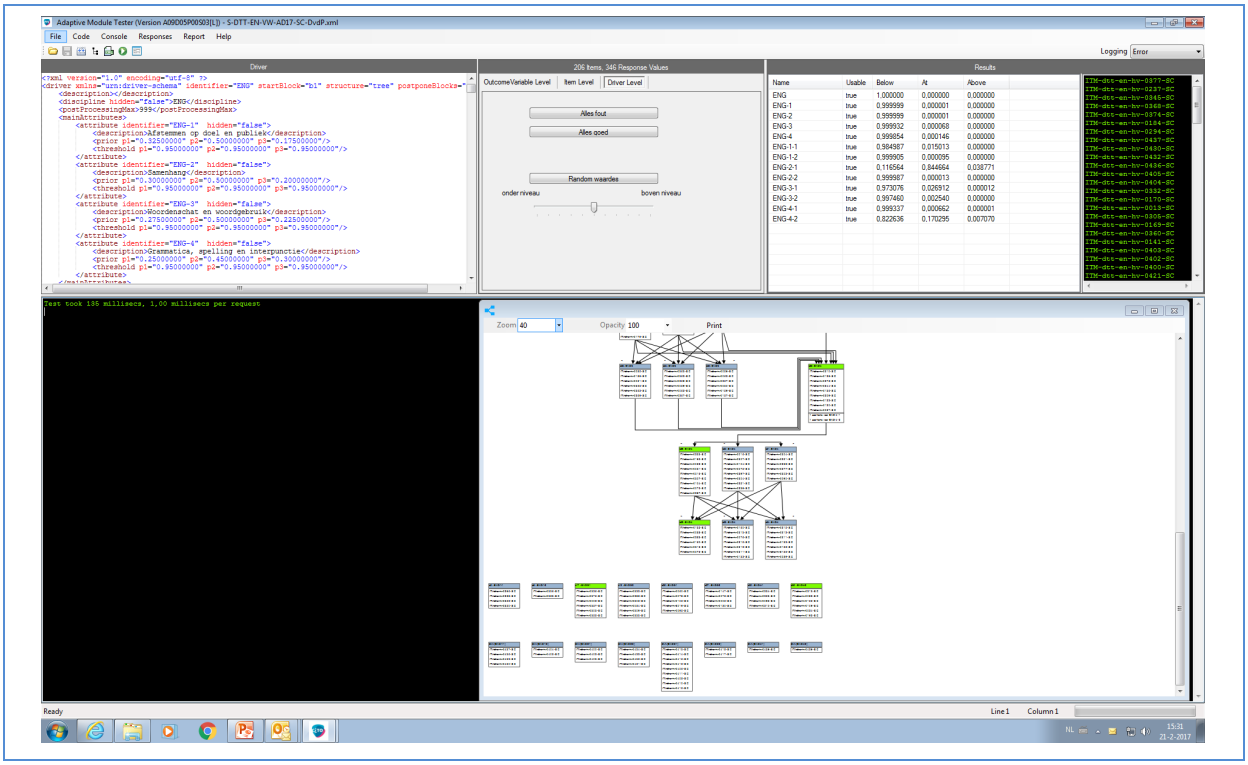


Figure 13-4. Screen print of the tool for testing the driver files

An adaptive test consists of a package including the tasks, the adaptive module, and a driver file. The driver file contains the settings for the adaptive test:

1. The parameters of the responses or combinations of responses for testlets (P<; P= and P>)
2. The starting values (prior model probabilities)
3. Settings for the adaptive rules
    a. The boundary values for exiting or skipping
    b. The usability setting (in principle, usable or not usable)
4. The adaptive design for the entire test: the block composition, the order of the main aspects, and the seed tasks (see Figure 13-5 for an example).

In the adaptive module, the probability of the diagnosis (the posterior model probability) is calculated after each task. A determination is also made concerning whether blocks may be skipped and whether the assessment outcome is usable. After the first stage, the further-testing algorithm is used to determine the order in which the sub-aspects will be subjected to further testing.

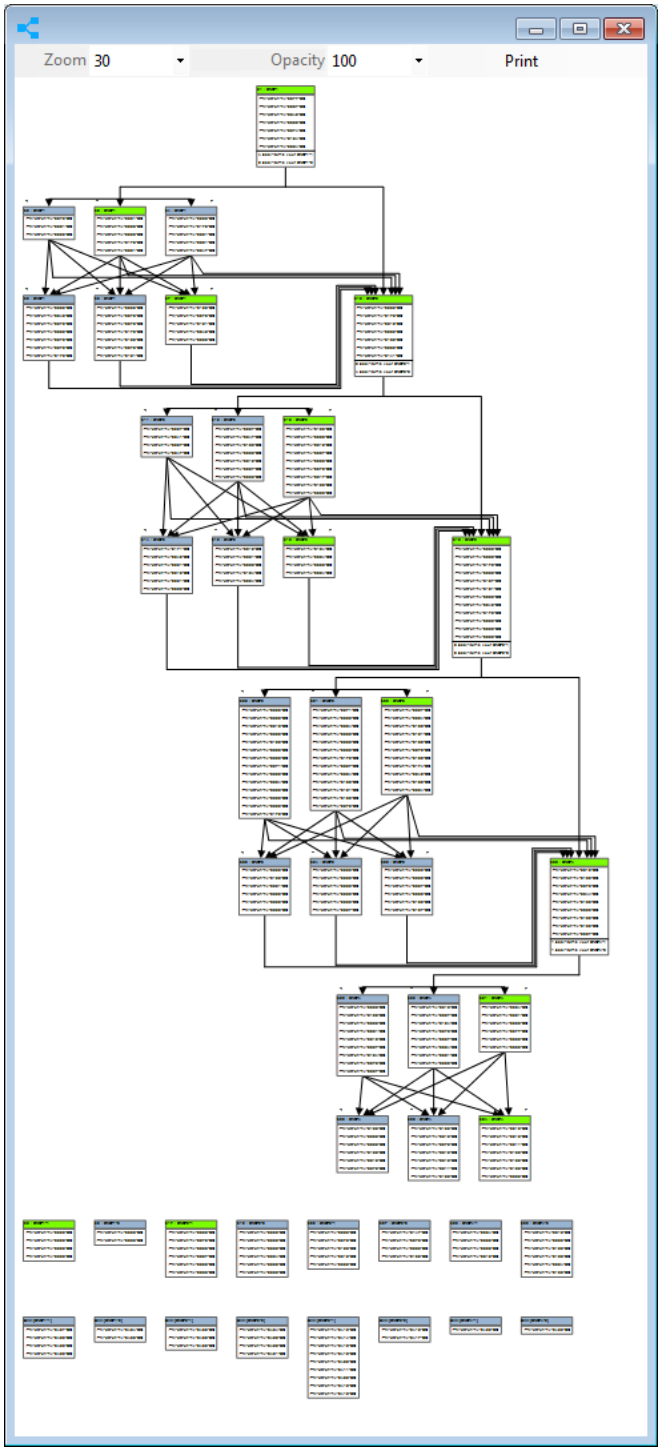


Figure 13-5. Visualization of an adaptive test, with the adaptive path of a student who has answered almost everything correctly indicated in green

13.8 Conclusion

For the DET, adaptivity is used in order to arrive at more diagnoses within the available administration time. Various grades of adaptivity are used in the DET. First, several different versions are used for each educational stream. Within each version, students' progress through two stages. In the first stage, diagnoses are made for all of the main aspects, and tasks are sown. In this stage, adaptivity occurs at the level of task blocks. In the second stage, the further-testing stage, diagnoses are made for the sub-aspects. The course of the test is regulated by three adaptive rules.

The adaptivity in the DET is based on the model developed by Sies and Hoijtink (2014). It is a completely new approach, and various aspects call for further investigation. The advantages and disadvantages of this approach could be investigated through an IRT-based approach (e.g., the approach proposed by Eggen and Straetmans, 2000). The impact of the uncertainty concerning the parameters on the accuracy of the diagnoses should also be examined further, as advised by Veldkamp (2012) for any operational computer-driven adaptive test. Further research should also be conducted on the adaptive rules, possibly allowing the development of alternative algorithms for further testing.

Adaptivity has increased during the development. In the first adaptive administration of 2016, adaptivity was used only in the first stage, in which diagnoses were made for the main aspects. At that time, only two layers of blocks were used. Although adaptivity is likely to increase even further in the future, a fully adaptive administration limits the possibilities for recalibration and quality control. Moreover, as concluded by Zwitser and Maris (2015), the introduction of additional stages in a multi-stage test does not necessarily increase efficiency. Further research is needed in order to identify a design that maximizes efficiency. One possibility for increasing efficiency could involve using the coherence between aspects or a possible hierarchical structure for adaptivity, as has been done by various scholars, including Gierl and Zhou (2008).

13.9 References

Cito (2015). *Diagnostische tussentijdse toets: Verslag pretest 2015 [Diagnostic Educational Test: Pre-test report 2015].* Arnhem: Cito.

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016 [Diagnostic Educational Test: Pre-test report 2016].* Arnhem: Cito.

Eggen, T. J. H. M, & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*, 713- 734.

Eggen, T. J. H. M. (2012). Computerized Adaptive Testing Item Selection in Computerized Adaptive Learning Systems. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC* (pp. 11-21). Enschede: RCEC, Cito/University of Twente.

Gierl, M. J., & Zhou, J. (2008). Computer adaptive attribute testing. A new approach to cognitive diagnostic assessment. *Zeitschrift fur Psychologie, 216*, 29-39.

Glas, C. A. W. (2000). Item calibration and parameter drift. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: theory and practice* (pp. 183–199). Dordrecht: Kluwer Academic Publishers.

Hendrickson, A. (2007). An NCME instructional module on multistage testing. *Educational Measurement: Issues and Practice, 26*(2), 44-52.

Hoijtink, H., & Schouwstra, S. (2014). Interpretatie van het initiële profiel op deelattribuutniveau [Interpretation of the initial profile at the level of sub-attributes]. In S. Schouwstra (Ed.), *De diagnostische tussentijdse toets: onderzoek 2014 [Diagnostic Educational Test: Research 2014]* (pp. 41-46). Arnhem: Cito.

Hoijtink, H., & Sies, A. (2013). De psychometrie van de diagnostische tussentijdse toets [The psychometrics of the Diagnostic Educational Test]. In S. Schouwstra (Ed.), *De diagnostische tussentijdse toets: onderzoek 2013 [Diagnostic Educational Test: Research 2013]* (pp. 13-62). Arnhem: Cito.

Hoijtink, H., & Sies, A. (2014). Adaptieve procedure met itemblokken [Adaptive procedure with item blocks]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [The Diagnostic Educational Test: Research 2014]* (pp. 33-40). Arnhem: Cito.

Hol, A. M. (2006). *A CAT with personality and attitude* [Phd Thesis]. Amsterdam: University of Amsterdam.

Maas, L. (2017). *Potential impact of item parameter drift on diagnostic accuracy in educational testing* [Master's Thesis]. Utrecht: Utrecht University.

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

Sies, A. (2014). *Cognitive Diagnostic Testing Using Calibrated Hypotheses* [Master's Thesis]. Utrecht: Utrecht University.

Veldkamp, B. P. (2012). Ensuring the future of computerized adaptive testing. In T. J. H. M. Eggen & B. P. Veldkamp (Eds.), *Psychometrics in Practice at RCEC* (pp. 35-46). Enschede: RCEC, Cito/University of Twente.

Zwitser, R. J., & Maris, G. K. J. (2015). Conditional statistical inference with multistage testing designs. *Psychometrika, 80*(1), 65-84.

14 The creation of a new form of reporting for the DET

Sanneke Schouwstra and Patrick de Klein

The DET is intended primarily to inform teachers and students on the student's strengths and points for improvement. The information provided by the DET is intended to equip teachers to adjust the learning processes of students based on their learning needs. To serve this formative function, it was necessary to pay careful attention to the reports. The importance of reporting was emphasized from the beginning of the development of the DET, and the report model was added to the conceptual framework of the assessment (see Chapter 1).

14.1 Stages in the development of the form of reporting

Four stages can be distinguished in the development of the DET reports. The blueprint for the report was worked out during the preliminary study. First, a literature study was conducted with regard to the principles and importance of feedback and reports. Several design principles that have been adopted in the creation of prototypical reports were derived from the literature study (see the example in Figure 14-1). For each subject, the prototypical reports were presented to teachers in field consultations.

The results of the field consultations revealed that several teachers had noted that they considered the information about reliability of the diagnoses less useful or difficult to interpret (Roelofs & Schouwstra, 2012). To ensure the fair and proper use of the results, it is crucial for the user to be provided with as much information as possible about the degree of certainty of the diagnoses. For this reason, the second stage included re-examining the visualization. It was in this stage that the 'dot view' was conceived (see Figure 14-2).

In the dot view, the most likely diagnosis for each aspect is indicated by a colored dot. For example, as displayed in Figure 14-2, the student's mastery of '*Coherence'* is probably at level. The other two diagnoses are displayed along with the most likely diagnosis. The more likely a diagnosis is, the larger and more visible the dot is. For example, although the diagnosis for '*Vocabulary and word usage'* is at level (blue dot), the student's answers suggest that the student might also be below level. The size of the orange dot indicates that this is reasonably likely.

In this second stage, sample student reports were created for each of the three subjects for the purpose of being submitted to the Assessment Specification committees for feedback. The student models and simulated data were used to create two student reports and two group reports:

- A student report for a student showing a point for improvement
- A student report for a student showing a strength
- A group report for a vmbo class
- A group report for a vwo class showing a specific point for improvement

In the third stage, a 'mock-up' was created, after the Assessment Specification committees had provided feedback. The mock-up was intended to provide an initial impression of how the DET report for teachers might look and work. Finally, in the last stage, the Executive Agency for the Department of Education [DUO] used this mock-up and other resources to realize the definitive report, under the direction of the Board of Tests and Examinations (in Dutch, *College voor Toetsen en Examens*, or CvTE). In this final stage, Cito created sample tasks to illustrate the reported diagnoses.

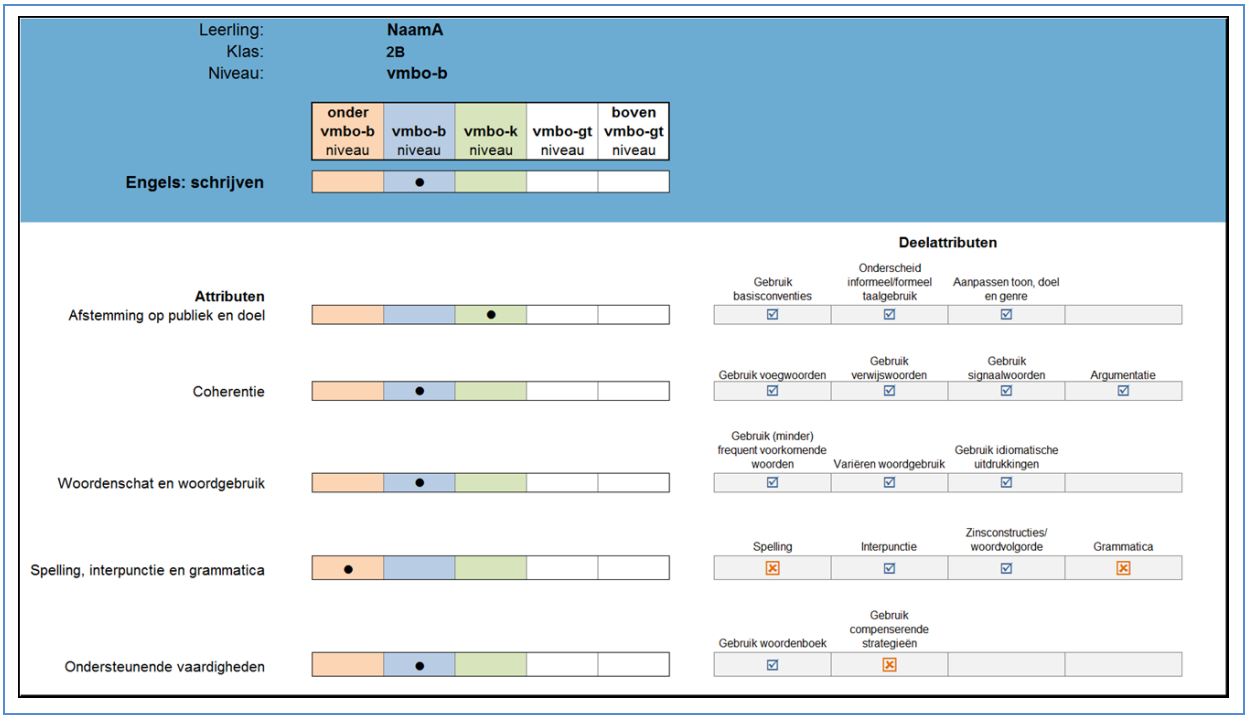


Figure 14-1. Example of a prototypical student report from the preliminary study (first stage) showing on the left side the diagnoses of the main aspects and on the right side of the sub-aspects for English writing

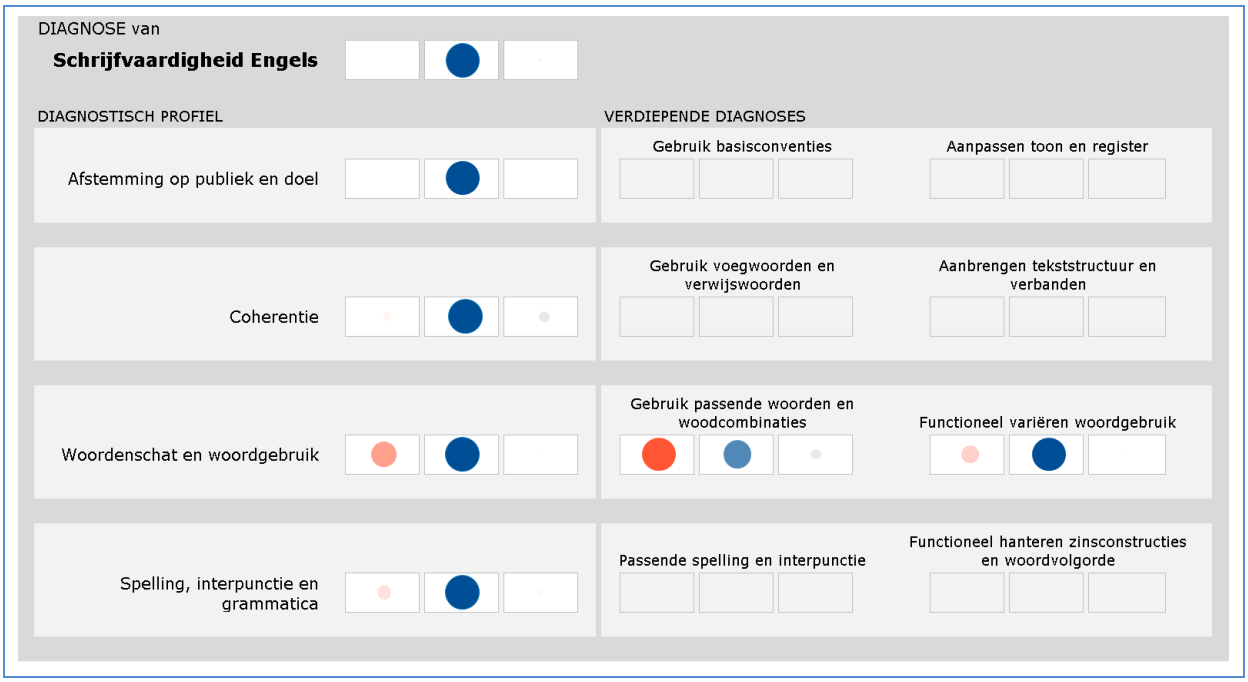


Figure 14-2. Example of a dot visualization (second stage) showing on the left side the diagnoses of the main aspects and on the right side of the sub-aspects for English writing

14.2 Basis for the DET report

The Explanatory Memorandum accompanying the legislative proposal Student-tracking system and the Diagnostic Educational Test (2013) provided for three levels of reporting: student, school and national. In the preliminary study, it was explained that group reports were highly advisable, given that the teacher is

the primary user. For this reason, four instead of three reports were planned during the development: one at the student level, one at school level, one at the national level and furthermore one at the group level. The national level was eliminated after the revised assignment on the part of the Ministry of Education, Culture and Science in 2014 (see Chapter 1).

All of the report designs were based on the following design principles (see also Roelofs & Schouwstra, 2012):

- Provide a clear image of the student's learning needs at a glance
- Communication through images (data visualization)
- Outcomes preferably in graphics, figures and text
- All information, images and colors should be used in a purposeful manner (i.e. do not use color for decorative purposes)
- Promote an intuitively correct interpretation of assessment outcomes
- Visualization of the certainty (or uncertainty) of the assessment outcomes
- Emphasis on remarkable outcomes
- Ease of use for teachers (e.g. simple navigation)

In this process, it was obviously ensured that the entire range of reports would be manageable, that the data would be grouped in a meaningful manner, that no statistical jargon would be used and that little text would appear on the screen. Another preference involved the creation of sample reports to accompany the reported assessment outcomes.

The reports for individual students should provide more detailed information than would be provided by a general performance or ability score. The reported outcomes are based on student models consisting of main aspects and sub-aspects: the DET diagnoses the aspects of knowledge and skills that are needed to write in English/write in Dutch/perform mathematics at the intended level. The student report was required to contain a visualization of this entire student model. For each aspect, the report contains three possible outcomes: the student's mastery is probably below level, at level or above level.

14.3 The mock-up

After the preliminary study and the static visualizations of the reports, a mock-up was created. The mock-up was developed to fulfil three objectives. First, it was created to serve as an illustration of a working report based on a student model, for use in presentations and training sessions (e.g. for the educational advisors who assisted the pilot schools in the implementation of the DET).[17] The mock-up was also needed in order to generate feedback from schools and teachers, so that the specifications for the definitive report could be improved. The Board of Tests and Examinations received two types of feedback on the report in the mock-up: through an online survey and through the advisory group (''klankbordgroep'' in Dutch) of the Board of Tests and Examinations. Third, the mock-up served to specify the realization in Facet by the Executive Agency for the Department of Education [DUO].

14.3.1 Development method for the mock-up

The mock-up was a working HTML5 prototype of the report module that was developed in an agile manner. Agile development is a rapid method that is very versatile and flexible. An entire cycle of design, implementation and evaluation is followed in short periods (known as iterations, see Figure 14-3). Each iteration results in a new version of the product (in our case, the mock-up).

The design principles from the preliminary study and the static visualizations (as in Figure 14-2) formed the base for the mock-up. The working prototype used data that had been simulated for previous psychometric research (Hoijtink & Sies, 2014). The simulated data were provided in an Excel file, from which they were converted to Json for use in Angular JS, an open-source web-application framework.

[17] The advisors were approached through Edventure, a professional association for educational consultancy firms.

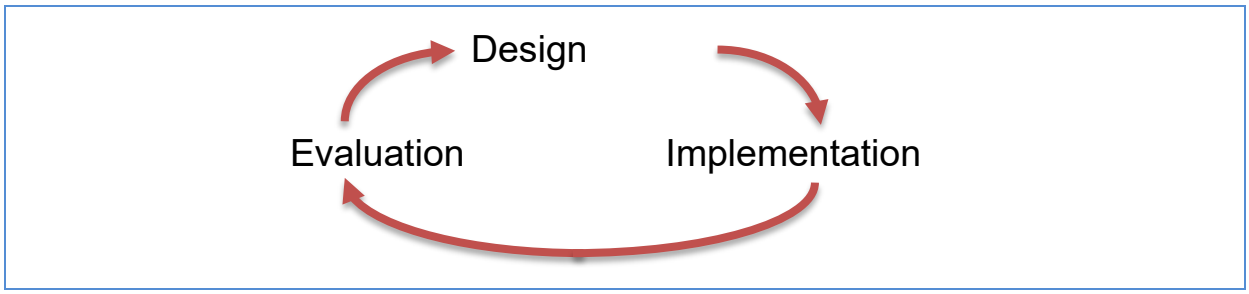


Figure 14-3. The cycle of one iteration

Reports were provided at four levels: the student level, the group level, the school level and the national level. The following users were specified during the preliminary study (Roelofs & Schouwstra, 2012):

- Student-level data: teacher, parent, student and mentor
- Group-level data: teacher, subject section, school management
- School-level data: school management
- National-level data: policymakers

The mock-up design process focused on the primary user: the teacher (see Chapter 1). For this reason, a student report and a group report were developed in the mock-up. At the beginning of the development of the mock-up, user stories were elaborated for teachers to use as a base for further development (for an example, see Figure 14 4). These user stories served as a base for the further development of the mock-up. The interactions and screens were elaborated in the report module based on the user objectives.

Who is the user?	The teacher
In which situation will the user use the report?	10-minute conference
What would the user like to achieve with the report?	To find and print a student report quickly, so that I can show it to the parents during the conference
What does this imply fort he design of the report?	A list from which a specific student can be selected
	Printable, preferably on one page
	Simple to use

Figure 14-4. Example of a user story

14.3.2 Design of the mock-up

A simple navigation between the starting screen, the student report and the group report were realized within the interaction design of the mock-up, as displayed in Figure 14-5. The most prominent characteristics of the static visualizations (see Figure 14-2) were incorporated into the visual design of the mock-up. First, the internal structure of main aspects and sub-aspects and all diagnoses for one student should be visible on a single screen. The text on the reports was restricted to a minimum.

One very important aspect that was incorporated in the visual design was the manner in which the level of certainty was visualized in dots. In the first form of the student report, only the most probable diagnosis was

displayed with a dot (the default view). In the second form, the other two diagnoses were displayed as well: the more likely the alternative diagnoses were, the larger and more visible the dots for the alternative diagnoses were (see Figure 14-6). A form was also created in which the size and visibility of all dots varied as a function of the likelihood of the diagnosis (the posterior model probabilities). In this form, the most probable diagnosis could also have a small dot. The plans also called for adding a numerical and a textual view of the diagnoses.

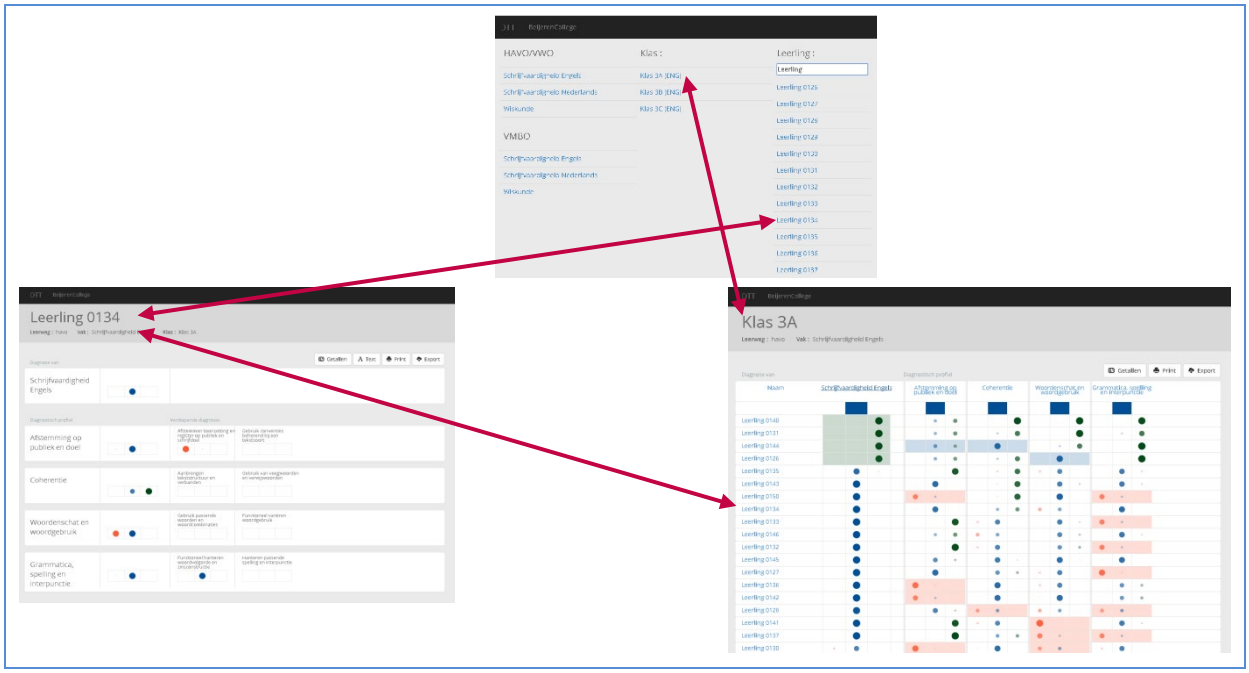


Figure 14-5. Interaction design

Figure 14-6. The three types of view

Other aspects of the definitive visual design included (see Figure 14-7):

- White field for content
- Clear interactivity using buttons
- Colors used only to indicate the meaning of the assessment outcomes
- The colors should result in the detection of important outcomes

The highlighting and options for sorting are worked out in the group report. For each class (group), the most common diagnoses in the class (or classes) for each aspect were highlighted (see Figure 14-8 at the top of each column). Remarkable outcomes were also highlighted for each student: students who needed additional attention (remedial or enrichment), as well as the aspects needing additional attention for each student.

With regard to sorting, the group report was required to allow sorting by name and the diagnosis for each main aspect. It also required convenient grouping for sorting by the overall diagnosis for the subject. This was realized by making an algorithm, in which within the sorting by overall diagnosis (below, at or above level), was sorted by the sub-aspect with the most below-level diagnoses within the group, and then by the sub-aspect with the second most below-level diagnoses and so forth (see the sorting in Figure 14-8).

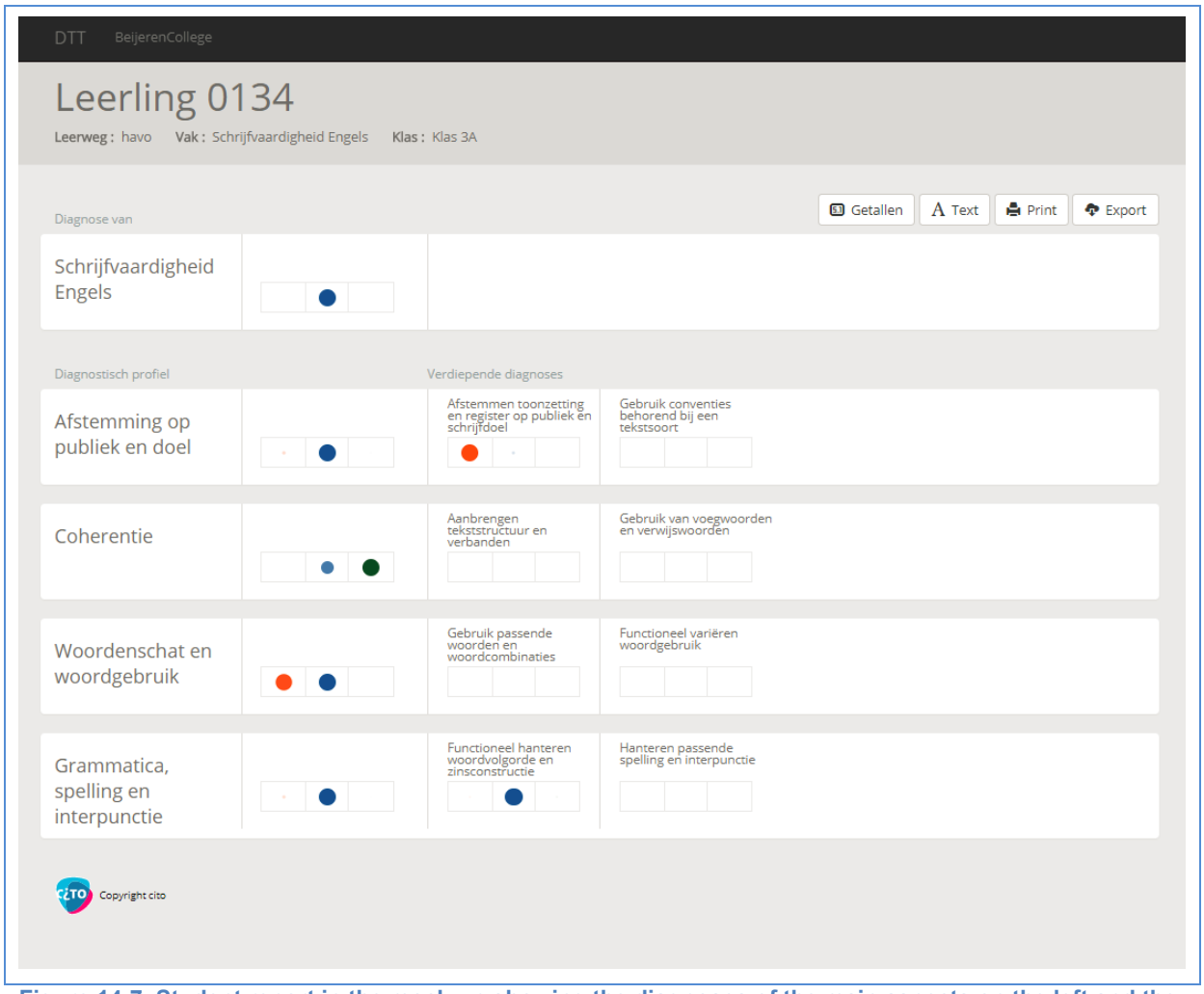


Figure 14-7. Student report in the mock-up showing the diagnoses of the main aspects on the left and the diagnoses of the sub-aspects on the right for writing English.

14.4 The definitive product

The final mock-up was used to gather feedback from the advisory group, as well as from schools and teachers. An online survey was administered to schools. According to the survey, the teachers and school directors found the visualization easy to understand, and they indicated that the student and class reports yielded information that was valuable and usable. It is interesting to note that the majority of the teachers and school directors primarily preferred to see the view that displayed the level of uncertainty. The graphic view with dots was also preferred to the numerical and textual views. The textual view was deemed the least necessary.

After the feedback was received, the report in Facet was realized under the direction of the Board of Tests and Examinations, using the mock-up as a specification. It was decided not to realize a textual description, as the feedback indicated that it would not be consulted very often. Based on the feedback, some changes were made to the highlighting in the group report. Because the assessment outcomes of the DET (without scores) are completely different from those for other tests in Facet, it was necessary to realize special adjustments in Facet.

The visualization of the assessment outcomes in the Facet report strongly resembled the visualization in the mock-up (see **Error! Reference source not found.**). The interactions were ultimately different than they had been in the design. For example, it was not possible for a teacher to view the reports on a tablet or

computer on site in a classroom (as described in the user stories). It was also not possible to download all of the reports for an administration at the same time.

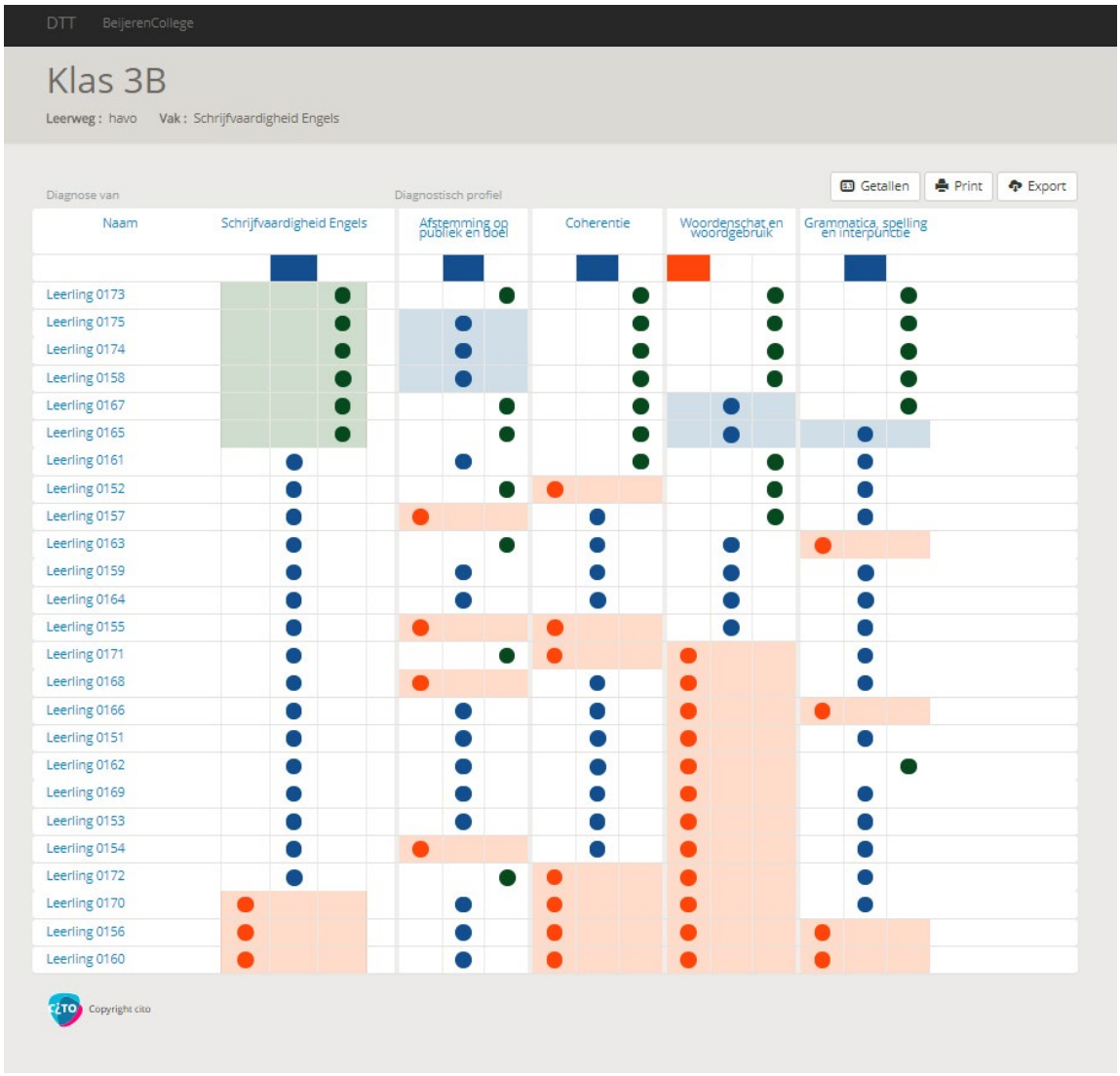


Figure 14-8. Group report in the mock-up, in the default view. On the left all students are listed and next to each student the diagnoses of the main aspects can be seen.

In addition to the student and group reports, the teachers wanted to view the completed tasks. In most cases, access to the tasks was used to serve three objectives. First, it is important to provide students with an idea in advance concerning the type of tasks that they would receive and, possibly, to allow them to practice with the new types of digital tasks and digital tools. It was also important for other parties (e.g. publishers) to see the type of tasks as well, so that they would be able to respond to developments and innovations. For this reason, practice tasks were created and posted on a website (https://oefenen.facet.onl). Second, it is important for each student to have the opportunity to verify that no errors have crept into the assessment. In case of doubt concerning the assessment outcome, therefore, it was possible to review the completed test under supervision. A third objective for which teachers commonly used the review option was to obtain a better grasp of the points for improvement and strengths emerging from the assessment outcomes, thus allowing adjustments to the teaching-learning process. To offer such links to teachers, it was decided to provide sample tasks along with the assessment outcome. If a teacher or student notices that the student has a specific point for improvement, they can work together to identify the types of tasks that had been difficult for the student in the DET.

For each sub-aspect, two sample tasks were developed, accompanied by brief descriptions of the diagnoses illustrated by the tasks, along with the key. These sample tasks could not yet be viewed directly through the student report, but they were available for teachers and students to review through a website. Ideally, teachers and students should be able to view the sample tasks directly through the report instead of through a website. Supplementary to the sample tasks, it was also possible to display an example of a teaching objective addressing the reported strengths or points for improvement.

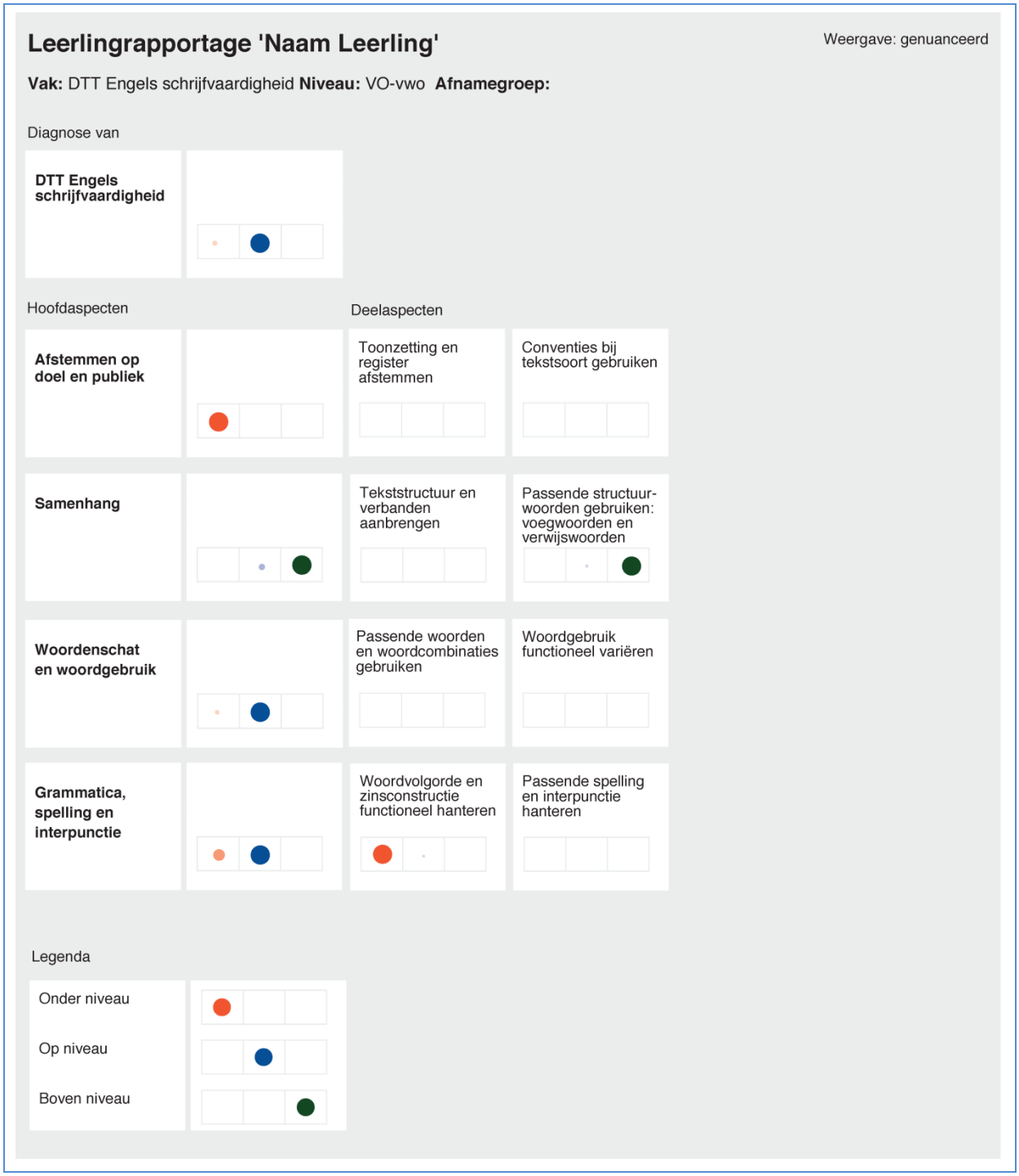


Figure 14-9. Example of a definitive student report, as it could be printed

14.5 References

Hoijtink, H., & Sies, A. (2014). Adaptieve procedure met itemblokken [Adaptive procedure with item blocks]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [The Diagnostic Educational Test: Research 2014]* (pp. 33-40). Arnhem: Cito.

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

15 Results of the 2017 DET administration

Sanneke Schouwstra, Jesse Koops, and Daniel van der Palm

The fully adaptive administration was held in February 2017. After the administration, the outcomes were analyzed. This analysis included the examination of the assessment paths, testing time, assessment outcomes, and the properties of the tasks (including seed tasks). New block arrangements for the adaptive tests were constructed with all of the approved tasks.

In 2017, 125 school locations participated in the administrations. In all, 30,490 administrations were started. This was fewer than had been started in 2016. Additional pre-test administrations in vmbo-bb were also conducted in 2017 (see Table 15-1), as too few observations had been realized in 2016 (<500) to arrive at a reliable calibration (Cito, 2016). The data from these pre-test administrations were merged with the data from the pre-test administrations in 2016 for recalibration.

Table 15-1. Number of administrations for each subject, by test

		English Writing Skills			Dutch Writing Skills			Mathematics		
		2015	2016	2017	2015	2016	2017	2015	2016	2017
Pre-test	vmbo-bb	505	358	213	747	383	273	618	343	316
	vmbo-kb	793	607		787	581		1117	753	
	vmbo-gt	1321	1475		1791	1397		1651	1586	
	havo	1269	1036		1321	820		1777	1314	
	vwo	1448	1150		1774	922		1839	1122	
	Total	5336	4626	213	6420	4103	273	7002	5118	316
Adaptive	vmbo-bb		808	598		988	809		919	729
	vmbo-kb		1266	1208		1362	1325		1574	1365
	vmbo-gt		2364	2319		3404	2955		2753	2396
	havo		2265	2681		1956	2366		3056	3211
	vwo		2110	2390		2255	2607		2013	2729
	Total		8813	9196		9965	10062		10315	10430
Total		5336	13439	9409	6420	14068	10335	7002	15433	10746

Comment: In each year, some students participated in multiple administrations (e.g., pre-test and adaptive) and/or in multiple subjects.

15.1 Paths taken in the adaptive administration

The adaptive character made it possible for students to proceed through the assessment in a wide variety of ways. The total number of possible paths for each assessment is astronomical. Even in the first stage (without further testing), there are 6561 different paths for each assessment. To provide some type of overview of the paths followed, the number of students with a given path through the various layers for each main aspect was visualized (see e.g., Figure 15-1).

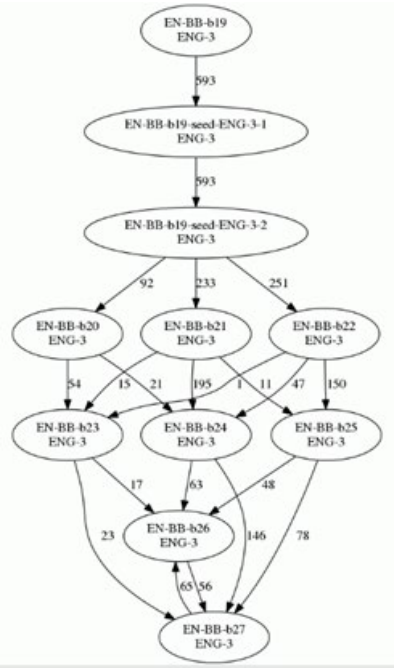


Figure 15-1. Example of a visualization of the paths for Main Aspect *3. Vocabulary and word usage,* English Writing Skills for vmbo-bb.

In the example, 593 students in vmbo-bb completed the seed tasks for the first Sub-aspect *3.1 Is capable of using words and combinations of words suited to the task* after the first layer of *3. Vocabulary and word usage* in Writing Skills. The same 593 students then completed the seed tasks for the second Sub-aspect *3.2 Is capable of functional variation in word usage.*

The first forwarding occurred after the seed tasks. In the second layer, the below-level block was presented to 92 students, the at-level block was presented to 233 students, and the above-level block was presented to 251 students. Students then started on the third layer.

- After the below-level block in the second layer, 54 students were also forwarded to the below-level block in the third level, and 21 students were forwarded to the at-level block. No students were forwarded to the above-level block.
- Of the students who had completed the at-level block in the second layer, 15 were forwarded to the below-level block in the third layer, 195 were forwarded to the at-level block, and 11 were forwarded to the above-level block.
- After the above-level block, one student was forwarded to the below-level block in the third layer, 47 were forwarded to the at-level block, and 150 were forwarded to the above-level block.

Further testing was started after the third layer.

- In all, 128 students started the further-testing block *3.1 Is capable of using words and combinations of words suited to the task*: 17 starting from the below-level block, 63 from the at-level block, and 48 from the above-level block. Of these students, 56 completed the further-testing block for the second sub-aspect.
- In all, 247 students started the further-testing block for the Sub-aspect *3.2 Is capable of functional variation in word usage,* and 65 of these students subsequently completed the further-testing block for the first sub-aspect.

Such visualizations were created for all three subjects, for each educational stream, and for each main aspect (for a total of 60; see Appendix 17.3). It was quite rare for students to be forwarded from a below-level block in the second layer to the above-level block in the third layer or, conversely, for students to be

forwarded from an above-level block in the second layer to the below-level block in the third layer. All of the other paths occurred regularly.

15.2 Testing time and response time

The average response time for a task in English Writing Skills was about 30 seconds. For Dutch Writing Skills, it was substantially longer: nearly 90 seconds. Response times for Mathematics varied considerably by educational stream. The average response time was about 60 seconds for vmbo, 90 seconds for havo, and nearly 120 seconds for vwo (see Table 15-2).

As indicated by the response times, several school locations used the option of allowing students to take a break. Because the data that were delivered contained no direct information about breaks, it must be derived from the response times. If, on average, the maximum response time for a given test across all students in a given administration far exceeded the average response time per task (e.g., longer than 20 minutes), it can be assumed that there was a break. In 7% of the administrations, the average maximum response time exceeded 20 minutes.

Although it was intended that students would spend three hours working on the DET, the data indicate that much less time was spent on the assessment (see Table 15-2). Correcting for breaks (total testing time minus the maximum response time in an assessment, averaged across all students), the data reveal that students spent more than an hour on the adaptive DET for English and 1.5-2 hours for Dutch and Mathematics.

Table 15-2. Average testing time and response time, by subject and educational stream

Subject	Educational stream	N	Average number of tasks	Average testing time	Corrected average testing time	Corrected average response time per task
English	vmbo-bb	598	139	00:58:51	00:54:57	00:00:25
	vmbo-kb	1208	151	01:11:33	01:07:29	00:00:28
	vmbo-gt	2319	149	01:09:37	01:06:35	00:00:28
	havo	2681	161	01:21:02	01:17:27	00:00:30
	vwo	2390	143	01:19:00	01:14:00	00:00:33
Dutch	vmbo-bb	809	81	01:39:02	01:29:00	00:01:13
	vmbo-kb	1325	93	02:02:46	01:50:01	00:01:19
	vmbo-gt	2955	90	02:10:26	02:00:36	00:01:27
	havo	2366	91	02:11:07	02:04:50	00:01:26
	vwo	2607	98	02:26:30	02:18:29	00:01:30
Mathematics	vmbo-bb	729	107	01:31:03	01:21:30	00:00:51
	vmbo-kb	1365	105	01:40:49	01:31:19	00:00:58
	vmbo-gt	2396	97	01:42:03	01:34:30	00:01:03
	havo	3211	71	01:45:27	01:38:10	00:01:29
	vwo	2729	74	02:21:03	02:09:53	00:01:54

15.3 Assessment outcomes

For each main aspect in each subject, the percentage of students receiving below-level, at-level, and above-level diagnoses was calculated (see Table 15-3, Table 15-4 and Table 15-5). The assessment outcomes were compared to the expected impact during the standard setting. The expected impact is the percentage of pre-test students falling below, at, or above level when the standard is applied (see also Chapter 11).

Table 15-3. Assessment outcomes for English writing in the 2017 adaptive administration, compared to the impact of the standards during the 2016 pre-test

		2017 assessment outcomes			Impact of standards for 2016 pre-test data		
English writing		% below level	% at level	% above level	% below level	% at level	% above level
vmbo-bb	1. Coordination with audience and objective	22.2%	55.4%	**22.4%**	13.5%	50.3%	36.2%
	2. Coherence	26.8%	**54.2%**	19.0%	28.3%	43.1%	28.6%
	3. Vocabulary and word usage	19.7%	45.5%	34.7%	12.9%	46.9%	40.2%
	4. Grammar, spelling, and punctuation	28.0%	45.0%	27.0%	24.2%	50.1%	25.6%
vmbo-kb	1. Coordination with audience and objective	20.1%	50.1%	29.8%	16.4%	44.4%	39.1%
	2. Coherence	25.0%	49.4%	**25.6%**	19.9%	42.4%	37.7%
	3. Vocabulary and word usage	18.3%	54.8%	26.9%	13.6%	50.9%	35.4%
	4. Grammar, spelling, and punctuation	15.2%	58.7%	26.0%	15.1%	67.1%	17.8%
vmbo-gt	1. Coordination with audience and objective	9.3%	60.1%	30.6%	12.8%	66.6%	20.6%
	2. Coherence	12.8%	47.4%	39.8%	15.1%	52.3%	32.6%
	3. Vocabulary and word usage	11.8%	**50.1%**	**38.1%**	11.1%	62.1%	26.8%
	4. Grammar, spelling, and punctuation	16.1%	64.5%	19.4%	13.2%	65.5%	21.3%
havo	1. Coordination with audience and objective	16.8%	62.8%	20.4%	13.1%	66.7%	20.2%
	2. Coherence	23.3%	54.3%	22.4%	26.7%	56.6%	16.8%
	3. Vocabulary and word usage	21.5%	54.9%	23.7%	23.6%	56.9%	19.5%
	4. Grammar, spelling, and punctuation	16.0%	59.4%	24.6%	19.1%	61.7%	19.2%
vwo	1. Coordination with audience and objective	22.0%	66.9%	**11.2%**	13.6%	64.0%	22.4%
	2. Coherence	**9.0%**	62.8%	28.2%	24.4%	54.9%	20.7%
	3. Vocabulary and word usage	**23.8%**	**52.1%**	24.0%	34.6%	40.4%	24.9%
	4. Grammar, spelling, and punctuation	**24.4%**	53.5%	22.2%	36.1%	46.8%	17.2%
Average		19.1%	55.1%	25.8%	19.4%	54.5%	26.1%

Comment: The percentages appearing in red are 10% lower than in the pre-test, and those appearing in green are 10% higher than in the pre-test.

Consistent with expectations, 55% of students were diagnosed as being at level for English Writing Skills, across all educational streams and aspects (see Table 15-3), with more than a fourth (26%) being above level and 19% below level. The most remarkable result is that, for three of the four main aspects in vwo, fewer students were diagnosed as being below level than had been the case during the pre-test.

Table 15-4. Assessment outcomes for Dutch writing in the 2017 adaptive administration, compared to the impact of the standards during the 2016 pre-test

Dutch writing		2017 assessment outcomes			Impact of standards for 2016 pre-test data		
		% below level	% at level	% above level	% below level	% at level	% above level
vmbo-bb	1 Rhetorical skills	11.0%	**51.5%**	**37.5%**	12.5%	62.7%	24.7%
	2 Text-structure skills	15.9%	57.7%	26.4%	24.1%	58.4%	17.5%
	3 Linguistic skills	21.1%	**61.8%**	17.1%	24.7%	49.8%	25.4%
	4 Orthographic skills	**23.7%**	**62.7%**	13.6%	36.7%	41.7%	21.6%
vmbo-kb	1 Rhetorical skills	10.1%	48.8%	41.1%	17.6%	41.4%	41.0%
	2 Text-structure skills	16.2%	50.9%	32.9%	17.9%	50.3%	31.8%
	3 Linguistic skills	9.2%	64.2%	26.6%	14.8%	54.7%	30.5%
	4 Orthographic skills	19.3%	66.3%	**14.4%**	14.5%	59.4%	26.1%
vmbo-gt	1 Rhetorical skills	4.5%	45.1%	50.4%	9.1%	47.7%	43.2%
	2 Text-structure skills	**7.2%**	64.5%	**28.3%**	28.6%	54.9%	16.5%
	3 Linguistic skills	7.9%	68.8%	23.3%	11.2%	62.8%	26.0%
	4 Orthographic skills	26.5%	**46.2%**	**27.4%**	20.0%	64.7%	15.3%
havo	1 Rhetorical skills	12.7%	59.8%	27.4%	12.9%	62.2%	24.9%
	2 Text-structure skills	18.5%	**60.5%**	20.9%	26.8%	48.9%	24.3%
	3 Linguistic skills	**17.3%**	**67.6%**	15.0%	35.0%	50.3%	14.7%
	4 Orthographic skills	**37.2%**	**59.8%**	**3.0%**	47.9%	38.8%	13.3%
vwo	1 Rhetorical skills	**1.4%**	70.1%	**28.6%**	15.2%	69.7%	15.2%
	2 Text-structure skills	13.5%	70.3%	16.2%	12.7%	70.5%	16.8%
	3 Linguistic skills	26.9%	61.2%	11.9%	18.4%	69.3%	12.3%
	4 Orthographic skills	23.9%	69.1%	7.0%	25.3%	62.7%	12.0%
Average		16.2%	60.3%	23.5%	21.3%	56.0%	22.7%

Comment: The percentages appearing in red are 10% lower than in the pre-test, and those appearing in green are 10% higher than in the pre-test.

For Dutch Writing Skills, 60% of the students are usually at level, with slightly less than a fourth being above level and 16% being below level (see Table 15-4). The percentages nevertheless vary considerably across aspects within the various educational streams. The most remarkable result is that more havo students were diagnosed as being at level (and fewer as being below level) than had been expected. For vmbo-bb, more students were diagnosed as being at level for Main Aspects *3 Linguistic skills* and *4 Orthographic skills*. For four aspects (in various educational streams), more students were diagnosed as being above level than had been expected.

Table 15-5. Assessment outcomes for mathematics from the 2017 adaptive administration, compared to the impact of the standards during the 2016 pre-test (domains B, C, and F) and the 2015 pre-test (domains D and E)

Mathematics		2017 assessment outcomes			Impact of standards during the 2016 pre-test*		
		% below level	% at level	% above level	% below level	% at level	% above level
vmbo-bb	B: Numbers (havo/vwo and variables)	43.2%	46.6%	10.2%	33.7%	53.2%	13.2%
	C: Relationships	**56.0%**	37.3%	6.8%	35.8%	45.7%	18.5%
	D: Measurement and geometry	32.7%	54.6%	12.7%	30.4%	54.0%	15.5%
	E: Associations and formulas	34.8%	53.8%	11.4%	27.1%	56.5%	16.4%
vmbo-kb	B: Numbers (havo/vwo and variables)	41.2%	46.3%	12.5%	36.6%	45.7%	17.7%
	C: Relationships	46.3%	43.7%	10.0%	42.0%	40.9%	17.1%
	D: Measurement and geometry	47.1%	49.4%	3.5%	37.5%	56.6%	5.9%
	E: Associations and formulas	**37.4%**	**53.4%**	9.2%	22.9%	65.4%	11.7%
vmbo-gt	B: Numbers (havo/vwo and variables)	35.6%	51.9%	12.5%	32.4%	53.7%	13.9%
	C: Relationships	36.2%	47.7%	16.1%	30.3%	53.3%	16.4%
	D: Measurement and geometry	33.8%	56.8%	9.4%	31.2%	62.6%	6.3%
	E: Associations and formulas	27.9%	**57.0%**	15.0%	23.4%	70.7%	5.9%
havo	B: Numbers (havo/vwo and variables)	**48.0%**	**46.3%**	5.6%	18.2%	66.4%	15.4%
	C: Relationships	**45.2%**	**42.5%**	12.4%	17.1%	66.1%	16.8%
	D: Measurement and geometry	37.4%	49.3%	13.3%	39.1%	49.3%	11.6%
	E: Associations and formulas	**46.6%**	**40.6%**	12.8%	28.2%	65.8%	6.0%
	F: Information processing and uncertainty (havo/vwo)	**54.8%**	**39.9%**	**5.4%**	16.4%	67.6%	16.0%
vwo	B: Numbers (havo/vwo and variables)	**42.2%**	**51.7%**	6.2%	11.4%	78.0%	10.6%
	C: Relationships	**45.7%**	**38.1%**	16.1%	17.2%	64.2%	18.5%
	D: Measurement and geometry	36.8%	53.7%	9.5%	31.8%	58.4%	9.8%
	E: Associations and formulas	47.5%	**46.8%**	5.7%	38.3%	57.7%	4.0%
	F: Information processing and uncertainty (havo/vwo)	**53.9%**	**41.4%**	**4.7%**	14.9%	69.7%	15.4%
Average		41.6%	48.2%	10.2%	27.3%	60.1%	12.6%

Comment: The percentages appearing in red are 10% lower than in the pre-test, and those appearing in green are 10% higher than in the pre-test.

For Mathematics (Table 15-5), in most cases, *fewer than half* of the students receive at-level diagnoses, with more than 40% receiving below-level diagnoses. The assessment outcomes are thus much lower than expected. The percentages of students falling below level are particularly high for havo and vwo.

A question that has not been studied yet, is whether the selected approach of adaptivity for the DET in Mathematics has made it too difficult. During the pre-tests, students received tasks from two (vmbo) or three (havo/vwo) domains. In the adaptive administration, students receive tasks from all four (vmbo) and all five (havo/vwo) domains. Moreover, in the pre-tests, all tasks were presented by domain (e.g., first, all tasks from domain *D: Measurement and geometry*, followed by all tasks from domain *E: Associations and formulas*). During the adaptive administration, the tasks from the various domains were presented in combination, in accordance with the recommendations of the confirmation committee (see Chapter 13). In mathematics, such alternation might make the assessment as a whole too difficult.

15.4 Task analyses and calibration (or re-calibration)

In 2017, a limited number of tasks were seeded. The seed tasks were analyzed and assessed in the same manner applied in the pre-test (see Chapter 4, Chapter, 5 and Chapter 7). The seed tasks that were approved by the confirmation committee were calibrated based on the assessment outcome (see Chapter 12). For English Writing Skills, 62 of the 75 seed tasks were approved. Of the 23 seed tasks for Dutch Writing Skills, 17 were approved, and 75 of the 96 seed tasks for Mathematics were approved.

All of the other tasks, which had previously been pre-tested, were re-calibrated. To ensure that the parameters were based on sufficient observations, the pre-test data for each task were considered in combination with the data from one adaptive administration. First, data were imputed for the data from the adaptive administration (see Chapter 12). The pre-test data were then merged with the imputed adaptive data. The tasks that had been pre-tested in 2015 were considered in combination with the adaptive administration of 2016. The tasks that had been pre-tested in 2016 were considered in combination with the adaptive administration of 2017. Finally, the tasks were calibrated according to category of mastery (below, at, or above level). For the pre-test students, diagnoses were determined for each main aspect by applying the standard (see Chapter 11). For the students of the adaptive administrations, the diagnoses were determined during the administration (the assessment outcome).

15.5 Block arrangement for the delivery in 2018

Dutch Writing Skills

Simulations were used to search for the block arrangement (see Chapter 12) that would yield the highest possible number of correct diagnoses. The simulations indicated that the best block arrangements for Dutch Writing Skills generated accurate diagnoses (see Table 15-6). According to the simulations, on average, the percentage of correct diagnoses is around 90%. The percentage was lower than 70% for only one diagnosis. In havo, for Main Aspect *1. Objective and audience,* 67.3% received the correct diagnosis, while 81.6% of the below-level students and 79.7% of the above-level students received correct diagnoses.

On average, the percentage of correct diagnoses on sub-aspects was good: 76%, but somewhat lower percentages (<70%) were common, particularly in vmbo-gt (15 of the 36 sub-aspect diagnoses) and vwo (12 of the 39 diagnoses), see Appendix 17.2. In havo and vwo, three sub-aspects were not sufficiently accurate for reporting. In other words, the percentage of correct diagnoses was lower than 60% for one or more of the diagnoses (below, at, or above level). This was the case in havo for Sub-aspect *2.1 Selecting text elements* and in vwo for Sub-aspects *1.1 Estimating the prior knowledge and information needs of readers* and *2.4 Presenting a standpoint and providing suitable arguments* (see Appendix 17.2).

English Writing Skills

The simulations indicated that the best block arrangement for English Writing Skills generated highly accurate diagnoses of the main aspects. The relative quality of the solutions is displayed in Table 15-7. For example, in the diagnosis of Main Aspect *3. Vocabulary and word usage* for vmbo-bb, 96.4% of the below-level students, 96.6% of the at-level students, and 92.3% of the above-level students were diagnosed correctly.

On average, the percentage of correct diagnoses for English Writing Skills was around 93%. In all cases, the percentage of correct diagnoses for the main aspects exceeded 85% (see Table 15-7). For the sub-aspects, the percentage of correct diagnoses was slightly lower, as these diagnoses were based on fewer responses. Even in this case, however, the average percentage was high: 83.4% correct diagnoses *before* further testing (see Appendix 17.2 for all outcomes at the sub-aspect level). The only cases in which the percentage of correct classifications was relatively low was for vwo, with regard to Sub-aspect *1.1 Is capable of coordinating tone and register to the audience and writing objective* and the at-level diagnosis for Sub-aspect *4.2 Is capable of using appropriate spelling and punctuation* (around 63%, see Appendix 17.2).

Table 15-6. Percentages of correct diagnoses for Dutch writing, number of tasks, and responses in the simulations

Educational stream	Main Aspect	Below level	At level	Above level	Number of tasks	Number of responses
vmbo-bb	1. Objective and audience	95.0%	94.9%	94.9%	43	160
	2. Structure	93.3%	87.3%	84.8%	51	129
	3. Word and sentence level	95.5%	95.5%	95.4%	47	157
	4. Spelling and punctuation	94.9%	94.4%	94.4%	40	213
vmbo-kb	1. Objective and audience	93.6%	93.5%	93.5%	44	190
	2. Structure	90.6%	90.6%	90.6%	56	118
	3. Word and sentence level	97.0%	96.8%	96.8%	50	192
	4. Spelling and punctuation	96.2%	95.7%	95.7%	37	212
vmbo-gt	1. Objective and audience	89.0%	89.0%	89.0%	43	170
	2. Structure	95.2%	94.5%	94.7%	64	127
	3. Word and sentence level	94.7%	94.5%	94.5%	43	163
	4. Spelling and punctuation	85.9%	78.9%	73.6%	41	212
havo	1. Objective and audience	81.6%	67.3%	79.7%	42	126
	2. Structure	80.5%	76.7%	72.7%	67	160
	3. Word and sentence level	92.1%	92.8%	93.3%	43	174
	4. Spelling and punctuation	95.0%	95.0%	95.0%	36	204
vwo	1. Objective and audience	96.4%	82.9%	71.9%	50	106
	2. Structure	76.6%	74.4%	71.4%	67	164
	3. Word and sentence level	91.6%	91.6%	91.6%	51	168
	4. Spelling and punctuation	97.8%	97.6%	97.6%	34	176

Table 15-7. Percentages of correct diagnoses for English writing, number of tasks, and responses in the simulations

Educational stream	Main Aspect	Below level	At level	Above level	Number of tasks	Number of responses
vmbo-bb	1. Coordination with audience and objective	90.3%	90.3%	90.3%	51	107
	2. Coherence	94.1%	94.1%	94.1%	61	80
	3. Vocabulary and word usage	96.4%	96.5%	96.6%	58	129
	4. Spelling, punctuation, and grammar	92.9%	92.8%	92.8%	65	70
vmbo-kb	1. Coordination with audience and objective	89.0%	89.0%	89.0%	52	124
	2. Coherence	95.4%	95.4%	95.4%	65	99
	3. Vocabulary and word usage	97.1%	97.1%	97.1%	68	147
	4. Spelling, punctuation, and grammar	96.0%	96.0%	96.0%	70	109
vmbo-gt	1. Coordination with audience and objective	95.6%	95.6%	95.6%	43	134
	2. Coherence	96.1%	96.1%	96.2%	73	96
	3. Vocabulary and word usage	96.0%	95.9%	96.0%	60	131
	4. Spelling, punctuation, and grammar	93.6%	93.6%	93.6%	70	99
havo	1. Coordination with audience and objective	93.2%	93.2%	93.2%	48	149
	2. Coherence	91.3%	91.3%	91.3%	54	84
	3. Vocabulary and word usage	94.6%	94.6%	94.7%	78	130
	4. Spelling, punctuation, and grammar	93.3%	93.3%	93.3%	73	109
vwo	1. Coordination with audience and objective	88.2%	88.2%	88.2%	39	100
	2. Coherence	92.5%	92.5%	92.5%	51	92
	3. Vocabulary and word usage	91.7%	91.7%	91.7%	72	114
	4. Spelling, punctuation, and grammar	87.6%	87.6%	87.6%	52	82

Mathematics

The best block arrangements for Mathematics generated reasonably accurate diagnoses (see Table 15-8) when all of the tasks were used in the first stage (and thus without further testing). On average, the

percentage of correct classifications was around 82%. The level accuracy was lower than was the case for the languages, possibly because the number of responses was also considerably lower in many cases.

In vmbo, the average level of accuracy was 84%. The level of accuracy was somewhat lower in havo and vwo (80%) than it was in vmbo (see Table 15-8). At the sub-aspect level, the percentage correct for the at-level diagnosis was much too low (<60% correct) in nearly all cases. On average, the percentage correct for the below-level diagnosis (across the five educational streams and three sub-aspects) was 70.6%; for the at-level diagnosis, the accuracy was 50.4%, with an accuracy of 74.1% for the above-level diagnosis.

In 2016 as well, the level of accuracy for the sub-aspects was too low (<60%). For this reason, in 2017, all of the tasks were used for the diagnosis at the level of main aspects (the first stage), and there were no further-testing blocks for the sub-aspects. In addition, no reporting was done at the sub-aspect level. In 2017 as well, there was no domain in which all three sub-aspects were diagnosed with sufficient accuracy within the educational streams. For this reason, all of the tasks were again included in the first stage, and there were no further-testing blocks. Given that the diagnoses at the level of sub-aspects are not sufficiently accurate, the recommendation remains that reporting sub-aspect diagnoses should be avoided (see Appendix 17.2).

Table 15-8. Percentages of correct diagnoses for Mathematics, number of tasks, and responses in the simulations

Educational stream	Main Aspect	Below level	At level	Above level	Number of tasks	Number of responses
vmbo-bb	B. Numbers	87.8%	87.8%	87.8%	70	90
	C. Relationships	86.2%	86.2%	86.2%	56	76
	D. Measurement and geometry	83.5%	83.5%	83.6%	49	77
	E. Associations and formulas	79.5%	79.5%	79.5%	43	49
vmbo-kb	B. Numbers	80.5%	80.5%	80.5%	57	84
	C. Relationships	81.4%	81.4%	81.4%	46	70
	D. Measurement and geometry	83.6%	83.6%	83.6%	55	70
	E. Associations and formulas	88.6%	88.6%	88.7%	41	63
vmbo-gt	B. Numbers	82.6%	82.6%	82.6%	53	70
	C. Relationships	83.3%	83.3%	83.3%	67	77
	D. Measurement and geometry	83.8%	83.8%	83.8%	52	63
	E. Associations and formulas	87.2%	87.2%	87.2%	46	61
havo	B. Numbers and variables	79.0%	79.0%	79.0%	34	69
	C. Relationships	80.0%	80.0%	80.0%	38	47
	D. Measurement and geometry	77.4%	77.3%	77.4%	37	49
	E. Associations and formulas	82.0%	82.0%	82.0%	36	56
	F. Information processing and uncertainty	76.7%	76.7%	76.7%	41	56
vwo	B. Numbers and variables	77.9%	77.9%	77.9%	38	67
	C. Relationships	77.4%	77.4%	77.4%	34	46
	D. Measurement and geometry	84.7%	84.7%	84.8%	37	47
	E. Associations and formulas	85.3%	85.2%	85.3%	42	48
	F. Information processing and uncertainty	75.4%	75.4%	75.4%	36	48

15.6 References

Cito (2016). *Diagnostische tussentijdse toets: Verslag pretest 2016 [Diagnostic Educational Test: Pre-test report 2016]. Arnhem*: Cito.

16 Concluding thoughts

Sanneke Schouwstra

The prototype of the DET was developed for three subjects in the 2012-2017 period, under the direction of the Board of Tests and Examinations and in cooperation with the field of education. For five educational streams, a fully adaptive assessment that could be administered digitally was created for Dutch Writing Skills, English Writing Skills, and Mathematics. In addition, item banks were developed for Dutch Reading Comprehension and English Reading Comprehension. This report describes the substantive development of the DET by Cito. Cito was responsible for the assessment construction, psychometric analyses, design of the reporting model, and the development of an adaptive module.

To respond to new political requirements and changing preferences of the field of education, it was necessary to develop the assessment more quickly and in a more flexible and interactive manner, together with schools. The entire process of developing the instrument was surrounded by research and innovations in several areas, including assessment content, psychometrics, and technology. Outcomes from research and knowledge acquired from sources including schools, stakeholders, market parties and education consultants through reports, workshops, and presentations were used throughout the development process.

The DET is based on new student models, which describe the aspects of knowledge and skills that are necessary to performance at specific levels. New, more authentic types of tasks were developed for the DET, and all of them are suitable for automated assessment. For example, in the language tests, students can select and correct words. In the mathematics test, students can insert formulas, draw graphs, and create geometric constructions in particular tasks, all of which can be assessed automatically. In the interest of efficiency in diagnosing, the test is adaptive, and an adaptive module was developed for this purpose. The form of adaptivity that was elaborated was focused on the ability to make the greatest possible number of accurate diagnoses within the available administration time. Instead of a traditional score, as often in the case of summative tests, the DET uses a new psychometric model to indicate the likelihood that a student is below level, at level, or above level for each aspect. A new form of reporting was developed for these assessment outcomes.

During the research and the development of the prototype, various options were suggested for the further optimization and development of the DET. For example, for the languages (among other subjects), it would be good to include open-ended writing tasks and to develop a marking aid (e.g., using automated assessment). Such open-ended writing tasks could also be used to investigate the relationship between the diagnoses on sub-aspects and the ability to write well.

For mathematics, further research is possible in order to develop a more refined student model that can be diagnosed accurately and that is easy for teachers to use, regardless of the teaching methods they have adopted. Further research could also investigate whether a different adaptive construction (e.g., by domain) for mathematics might make the assessment easier for students. The current form of the assessment and the tasks appears to be quite difficult for students.

The adaptivity and reports could also be developed further in order to improve their ability to respond to the needs of schools and teachers. For example, the sample tasks could be directly linked through the student report. In supplement to these links, examples could be displayed of learning objectives and a lesson that address the strengths and points for improvement that have been reported. In additions, alternative adaptive algorithms could be developed. For example, the developed further-testing algorithm emphasizes points for improvement, as addressing these points is expected to enhance higher performance. It is nevertheless conceivable that teachers might prefer to have equal emphasis on points for improvement and strengths in the further tests (Hoijtink & Schouwstra, 2014). This would require the development of an alternative further-testing algorithm.

The further development of adaptivity could also be directed toward the reduction of testing time. Results of the evaluation conducted by the Board of Tests and Examinations (to appear) indicate that this is a clear

wish on the part of schools. Even during the preliminary study, it was suggested that the mastery of one main aspect could have implications for the mastery of another main aspect, as is the case when the main aspects are ordered in a hierarchy (Roelofs & Schouwstra, 2012, p. 109). For example, good mastery of linguistic skills might be a pre-requisite for the ability to apply structure to a text. In this context, students who have not mastered linguistic skills would be highly unlikely to have a mastery of structure. This could lead to a reduction in the number of tasks on structure for students who have not mastered linguistic skills. This calls for further scientific research on the internal cohesion of the aspects of the student models.

Another way to reduce the number of tasks involves refreshing tasks. According to a study by Hoijtink and Sies (2013; 2014b), the discriminating power of tasks is strongly determined by the number of tasks needed. The number of tasks needed decreases as discriminating power increases. When tasks are refreshed, the choice could be made to replace the tasks with lower discriminating power. In time, this would result in a item bank with greater discriminating power, and fewer tasks will be needed in order to make an accurate diagnosis.

Examples could include the incorporation of more layers in the adaptive procedure or even the implementation of adaptivity at the task level. As indicated by previous research (Hoijtink & Sies, 2014), however, the block procedure was almost as accurate as adaptivity at the task level. The time gained at the same level of accuracy would thus be minimal. Although testing time could obviously be reduced by allowing less accurate diagnoses, the large number of incorrect diagnoses would substantially decrease the usefulness of the DET.

Another possibility that might not reduce testing time very much but that would reduce test pressure would be to spread the assessment over two days (e.g., by doing the further testing on the second day). Other possibilities could include a set-up in which the teacher uses the results to determine whether and for whom further testing will be done. Another advantage of having a clear distinction between the first stage and the further-testing stage is that it would be clear when sufficient tasks have been completed in order to prepare a report on the diagnosis of the main aspect. Such a clear indication is important in order to ensure that schools do not have students discontinue the assessment too soon, before sufficient tasks have been completed for arriving at a diagnosis.

There are obviously other possibilities for ensuring that it is clear when sufficient tasks have been completed for arriving at a diagnosis of the main aspects. For example, an information screen could be displayed during the assessment to notify students when they have completed enough tasks for a diagnosis. It is nevertheless crucial for schools and teachers to be well informed with regard to the necessary testing time and the possible consequences of allowing insufficient time for the DET (e.g., no report).

An obvious area for growth is the expansion of the assessment. The tasks for English Reading Comprehension and Dutch Reading Comprehension could be pre-tested. In addition, tasks could be developed for a diagnostic measurement of English listening skills, Dutch listening skills, and possibly arithmetic. These new skills might also call for the development of new types of tasks and the expansion of automated assessment. Another expansion that schools might need could involve a remedial test for determining whether students have benefitted from remedial lessons (if applicable), as well as a special version of the assessment for students with disabilities.

It can be concluded that the DET is a completely new product that offers many favorable opportunities for development and growth. It is an innovative assessment for all students at the end of the lower secondary education, based on student models, which is administered in an adaptive manner, includes new types of tasks, and features a new underlying psychometric model with a new form of reporting. The assessment identifies students' strengths and points for improvement, thus making it possible for teachers to take action for improvement and customization (e.g., using the series of lessons developed by Netherlands Institute of Curriculum Development [SLO, 2017]). The schools reacted positively to the new types of tasks and new forms of reporting, and they perceived the DET as a stimulus for formative evaluation (College voor Toetsen en Examens, to appear).

In June 2016, a political decision was taken to transfer the DET to market parties at the end of the pilot period. Several knowledge-exchange sessions for market parties have been held. In late May 2017 two market parties have announced they are going to offer the DET to schools, and in June the process of delivering the prototype to the Ministry of Education, Culture, and Science was started. The Ministry of Education, Culture, and Science has taken care of the disclosure and thus the transfer to the market.

16.1 References

College voor Toetsen en Examens (2017). *Slo-materiaal [Slo material]*. Retrieved december 2017, from https://www.pilotdtt.nl/slo-materiaal.

College voor Toetsen en Examens (forthcoming). *Eindrapport project DTT [Final report project DTT]*. Utrecht: College voor Toetsen en Examens.

Hoijtink, H., & Schouwstra, S. (2014). Interpretatie van het initiële profiel op deelattribuutniveau [Interpretation of the initial profile at the level of sub-attributes]. In S. Schouwstra (Ed.), *De diagnostische tussentijdse toets: onderzoek 2014 [Diagnostic Educational Test: Research 2014]* (pp. 41-46). Arnhem: Cito.

Hoijtink, H., & Sies, A. (2013). De psychometrie van de diagnostische tussentijdse toets [The psychometrics of the Diagnostic Educational Test]. In S. Schouwstra (Ed.), *De diagnostische tussentijdse toets: onderzoek 2013 [Diagnostic Educational Test: Research 2013]* (pp. 13-62). Arnhem: Cito.

Hoijtink, H., & Sies, A. (2014). Adaptieve procedure met itemblokken [Adaptive procedure with item blocks]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [The Diagnostic Educational Test: Research 2014] (*pp. 33-40). Arnhem: Cito.

Hoijtink, H., & Sies, A. (2014b). Een reconstructie van de succeskansen in de eerste try-out [A reconstruction of the success probabilities of the first try-out]. In S. Schouwstra (Ed.). *De diagnostisch tussentijdse toets: onderzoek 2014 [The Diagnostic Educational Test: Research 2014]* (pp. 11-32). Arnhem: Cito.

Roelofs, E., & Schouwstra, S. (Eds.) (2012). *Diagnostische tussentijdse toets: Verslag van de voorstudie [Diagnostic Educational Test: Results of the preliminary study].* Arnhem: Cito.

17 Appendices

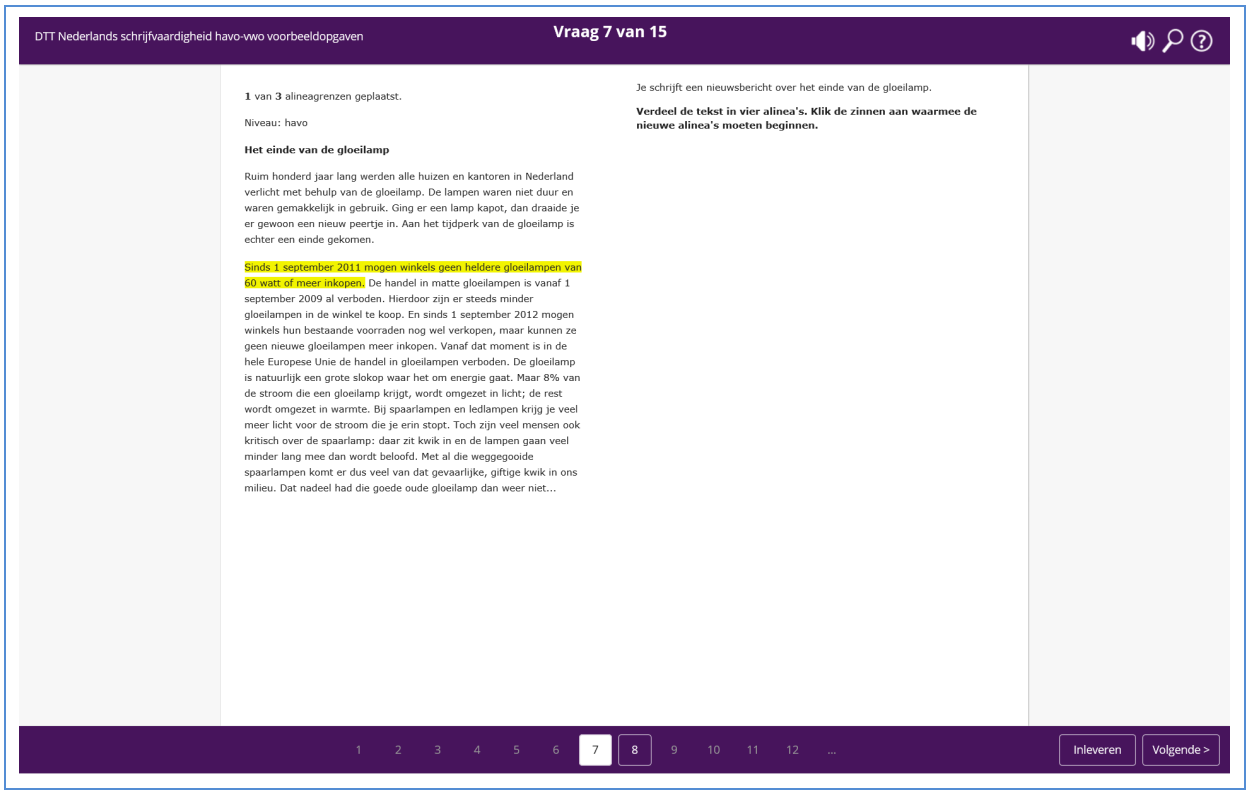
17.1 Appendix Examples of task types

All examples of tasks are in Dutch.

Paragraph task

A type of task in which the student divides a text into paragraphs. The paragraph divisions are visible in the text. Suitable for allowing the student to apply text structure.

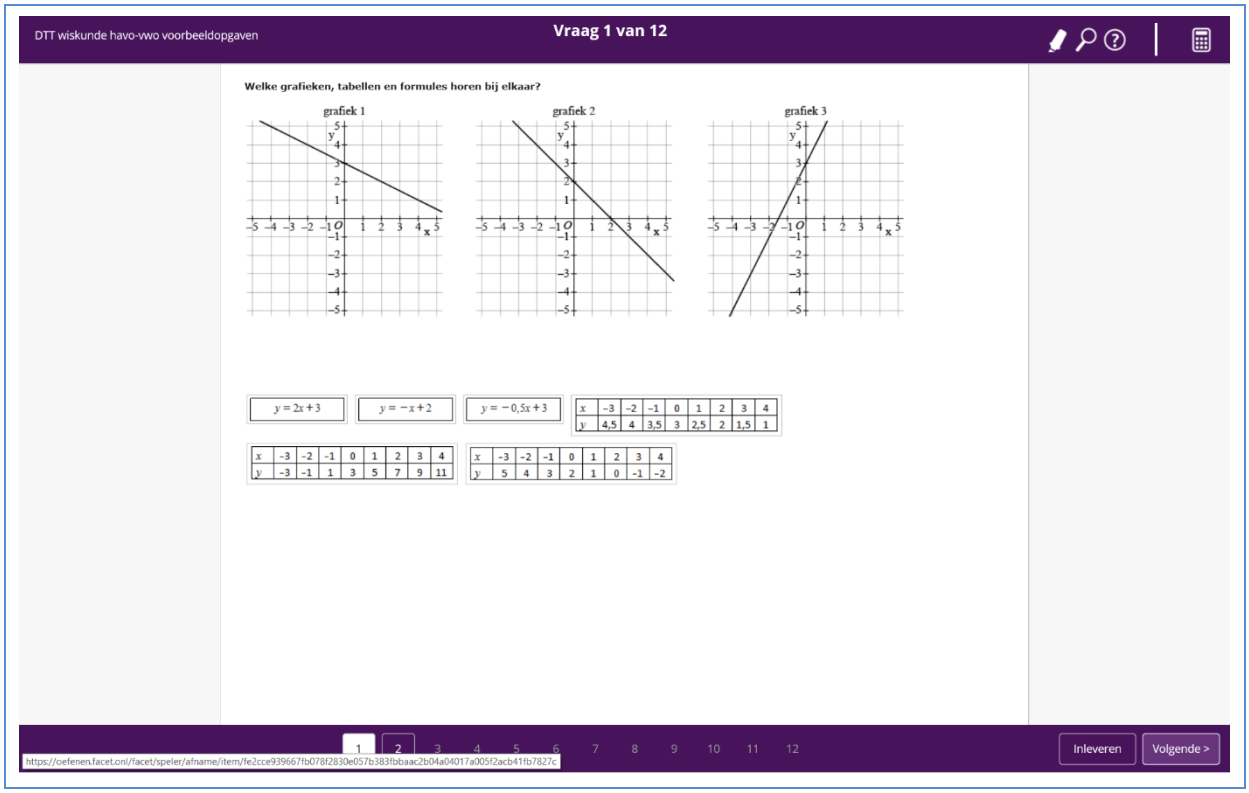
Kind of interaction: Clicking, multiple choice



Categorization task

The student drags particular elements (text, images) in fields by category. Suitable for sorting elements (e.g., formal/informal for tasks concerning types of text).

Kind of interaction: Dragging



Combination task

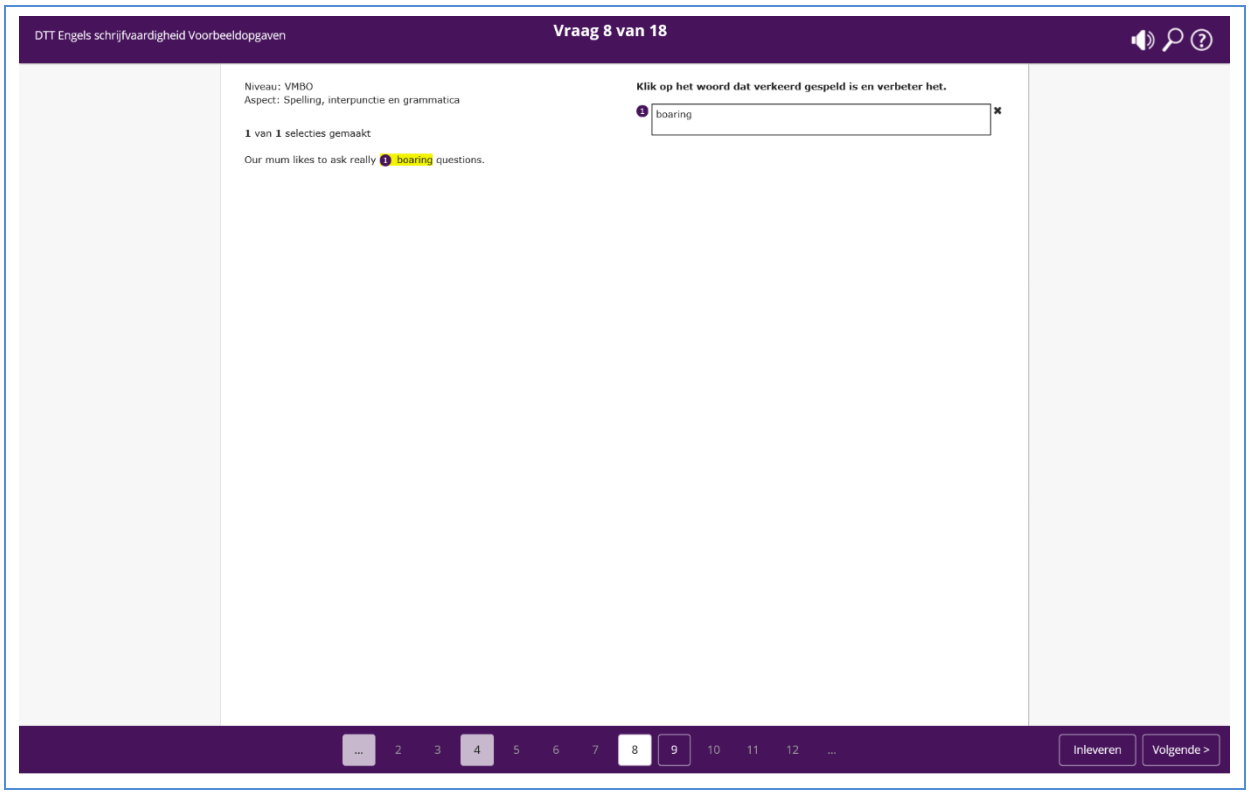
A type of task in which multiple types of interactions (e.g., multiple choice and open-ended task) are combined on a single screen.

Kind of interaction: Miscellaneous

Correction task

The student selects and improves components of a text. Suitable for assessing grammar and spelling.

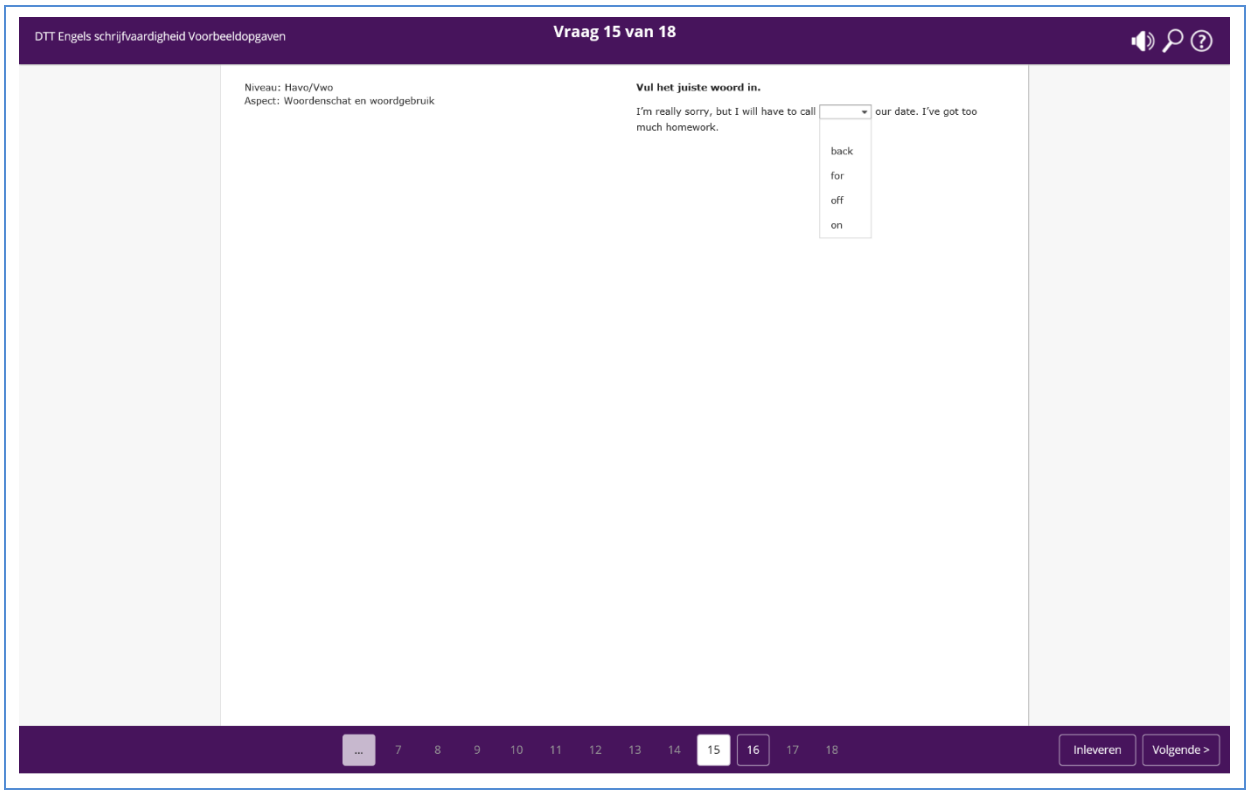
Clicking and entering



Drop-down task

A type of task in which the student selects an answer from a list of options. The student immediately sees the effect of the choice in the text. This type of task is often used for vocabulary questions in which the student inserts the most suitable word into the text.

Kind of interaction: Clicking, multiple choice



DTT Nederlands schrijfvaardigheid vmbo voorbeeldopgaven

Vraag 6 van 12

Niveau: KB

Je hebt een bericht geschreven over een opmerkelijke gebeurtenis. In de tekst ontbreekt een aantal woorden.

Kies het juiste woord om een logische tekst te krijgen.

Eeneiige tweelingbroers getrouwd met eeneiige tweelingzussen

In India zijn afgelopen donderdag twee tweelingbroers getrouwd met twee tweelingzussen. De twee broers waren hier blij mee, [ ] ze al van jongs af aan met tweelingzussen wilden trouwen. Ze hebben vijf jaar gezocht naar de perfecte vrouwelijke tweeling en hebben oproepen gedaan in kranten en op websites.

Volgens een van de bruidegoms was het een perfecte dag, [ ] hij het jammer vond dat er in het publiek niet veel tweelingen zaten. [ ] is er nog meer goed nieuws te melden: de vrouwen zijn alle een tweeling.

- Bovendien
- Daardoor
- Daarom

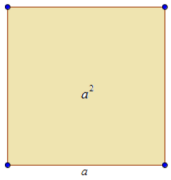
1 2 3 4 5 6 7 8 9 10 11 12

Inleveren | Volgende >

DTT wiskunde havo-vwo voorbeeldopgaven

Vraag 9 van 12

Het vierkant heeft zijde a en oppervlakte a^2.



Kies bij de gaten in de tekst steeds de juiste optie.

Als alle zijden van het vierkant twee keer zo groot worden, wordt de oppervlakte [ ] keer zo groot.

Om de oppervlakte van het vierkant te verdubbelen, moeten alle zijden vermenigvuldigd worden met [ ].

Als de oppervlakte van het vierkant drie keer zo groot is geworden, dan hebben de zijden een lengte van [ ].

- $1\frac{1}{2} \cdot a$
- $\sqrt{3} \cdot a$
- $3a$

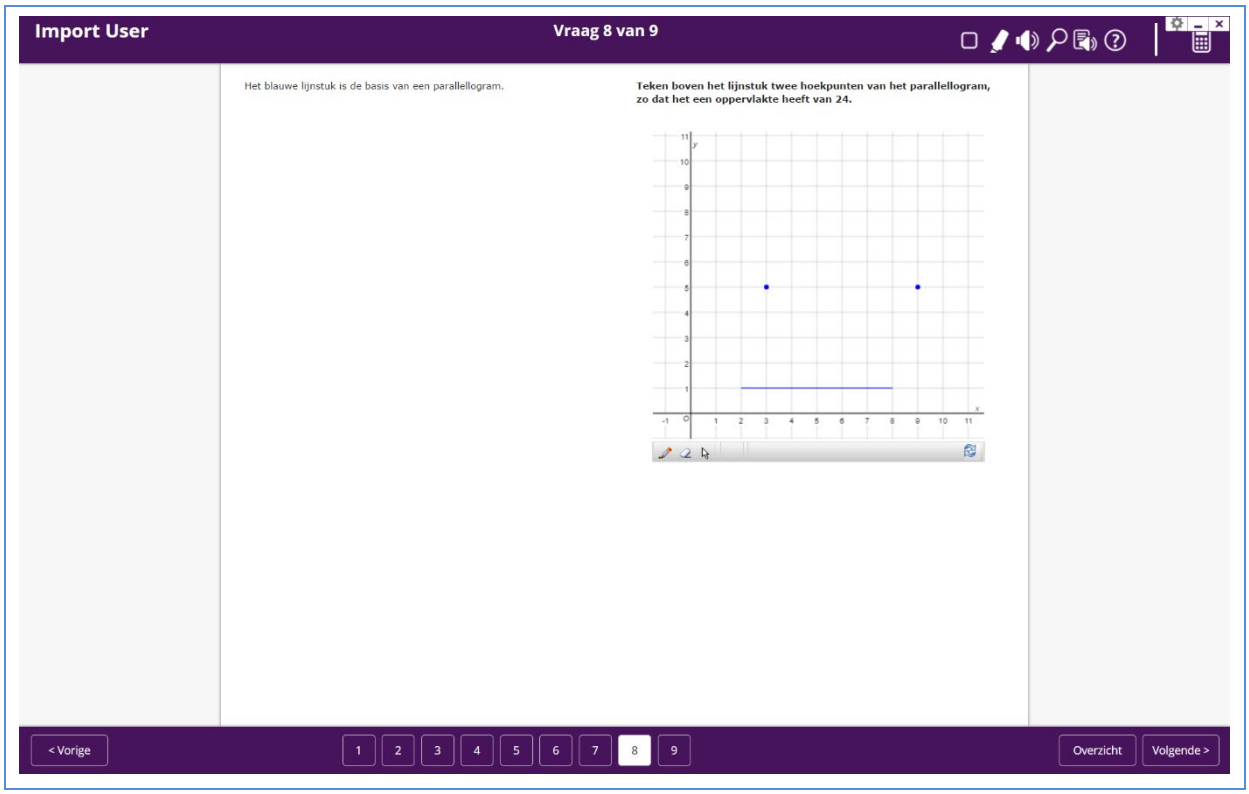
1 2 3 4 5 6 7 8 9 10 11 12

Inleveren | Volgende >

DME task (task using the Digital Mathematics Environment)

The student can draw chart points or charts as an answer.

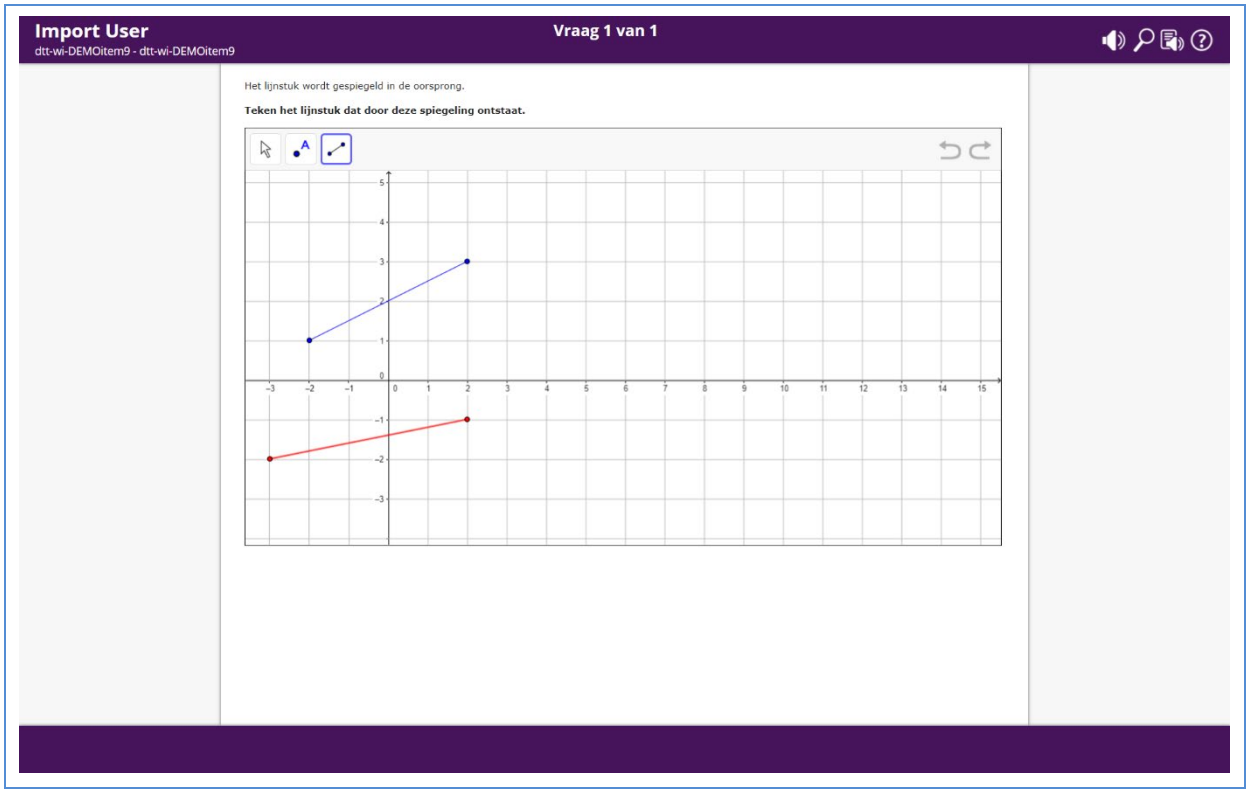
Kind of interaction: Miscellaneous



GeoGebra task

The student can perform various mathematical interactions as an answer. Example: drawing or adjusting the correct geometric figure.

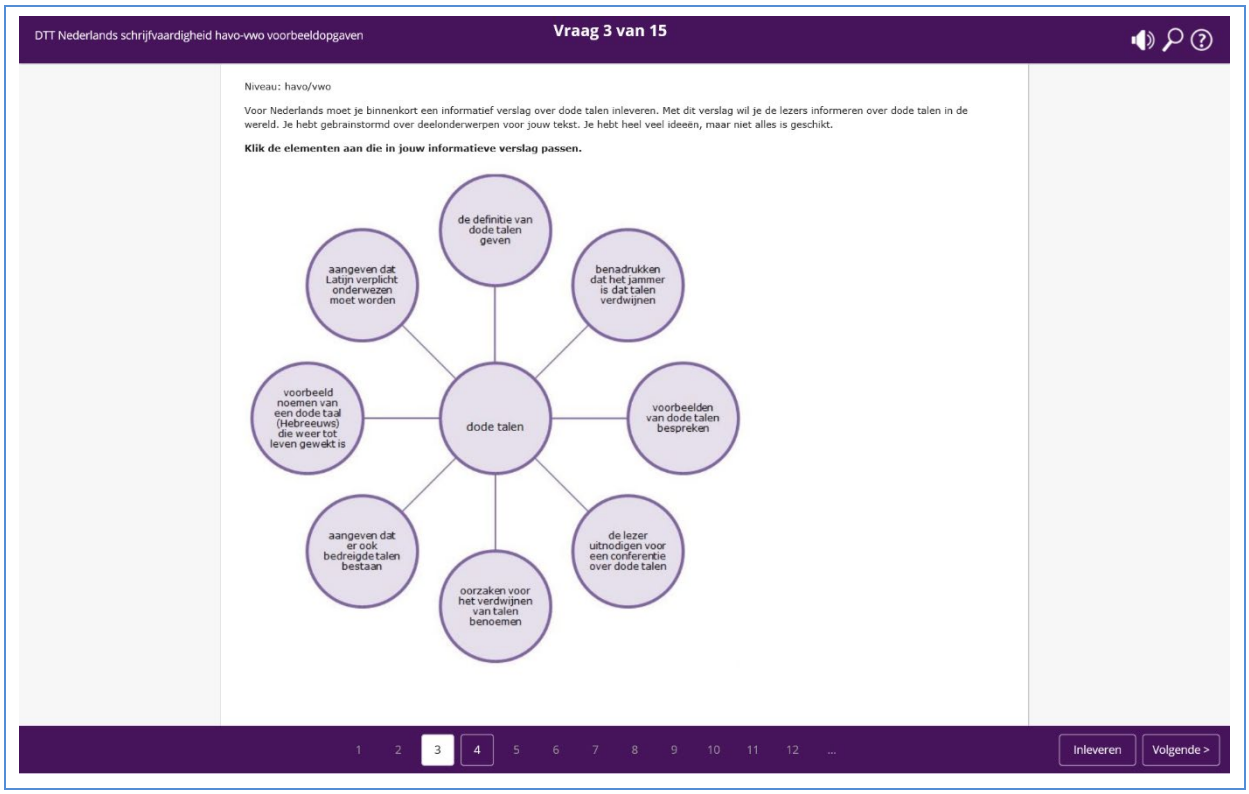
Kind of interaction: Miscellaneous

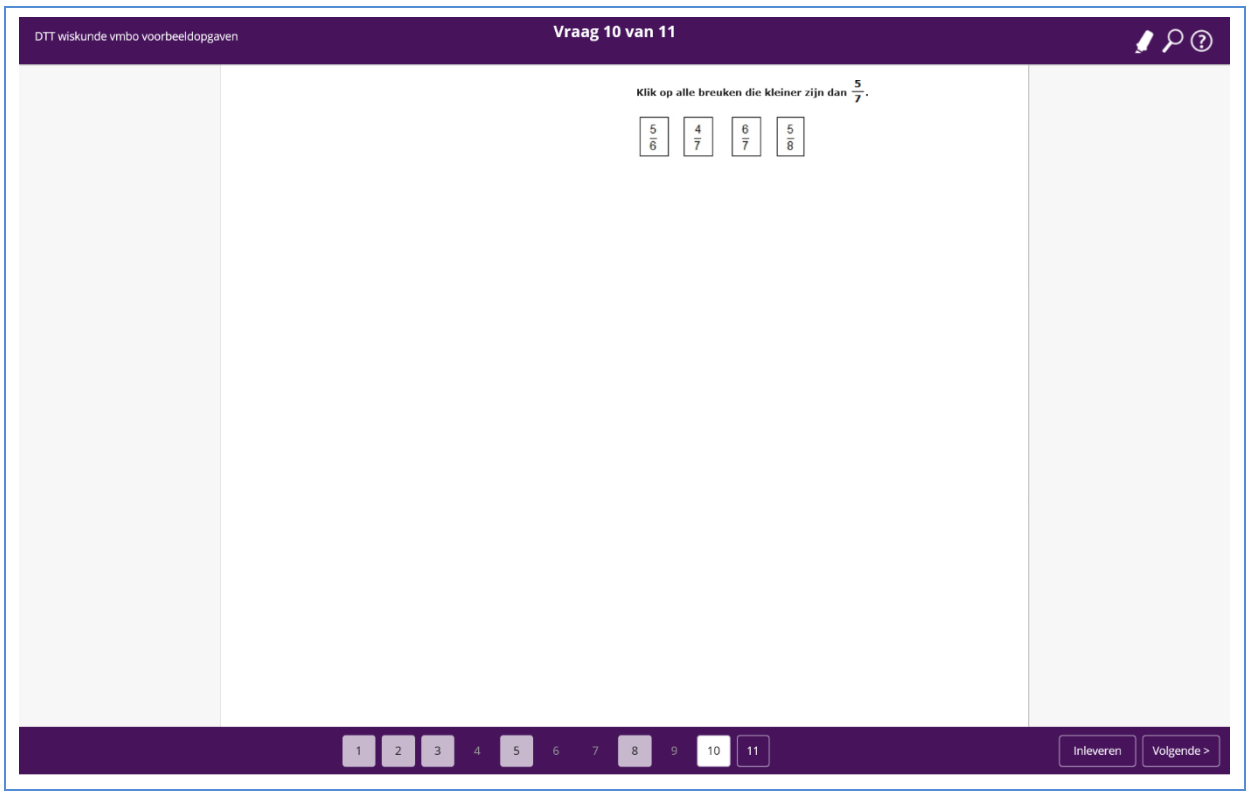


Hotspot task

The student clicks on the correct answer in an image.

Kind of interaction: Clicking

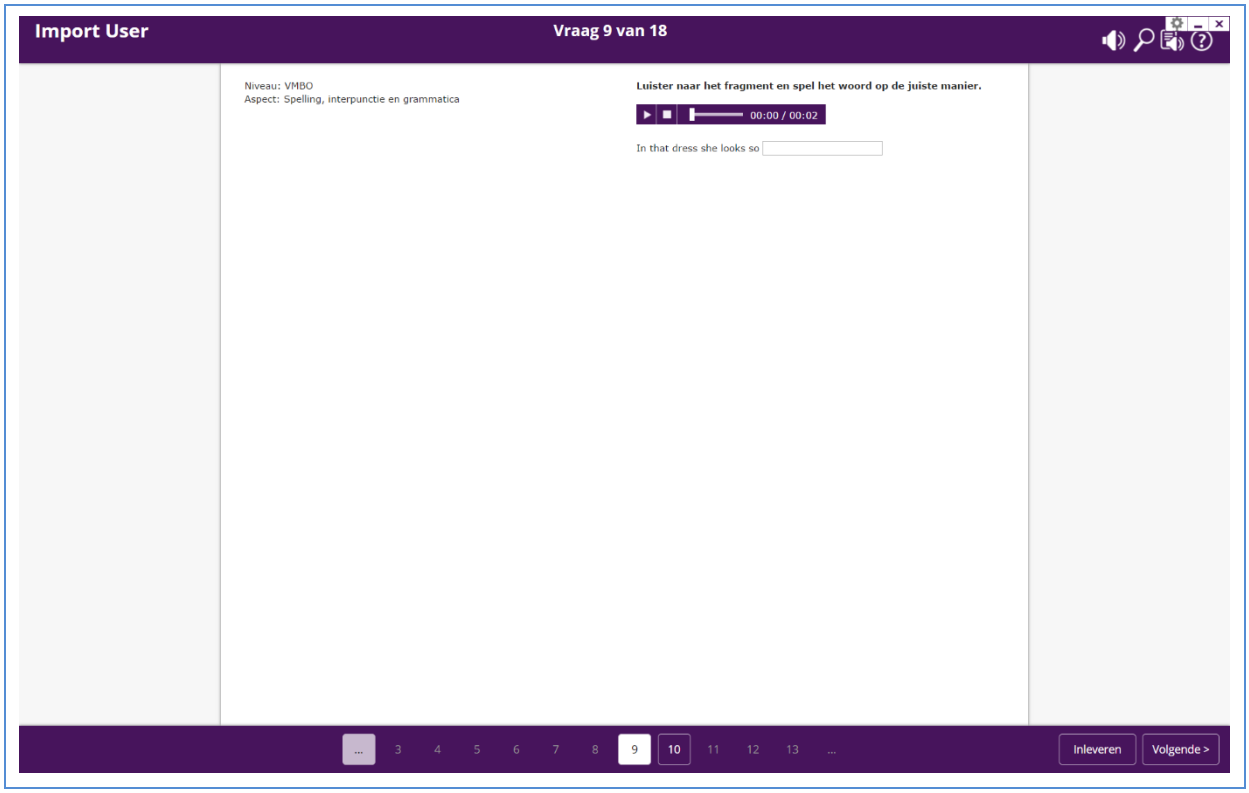


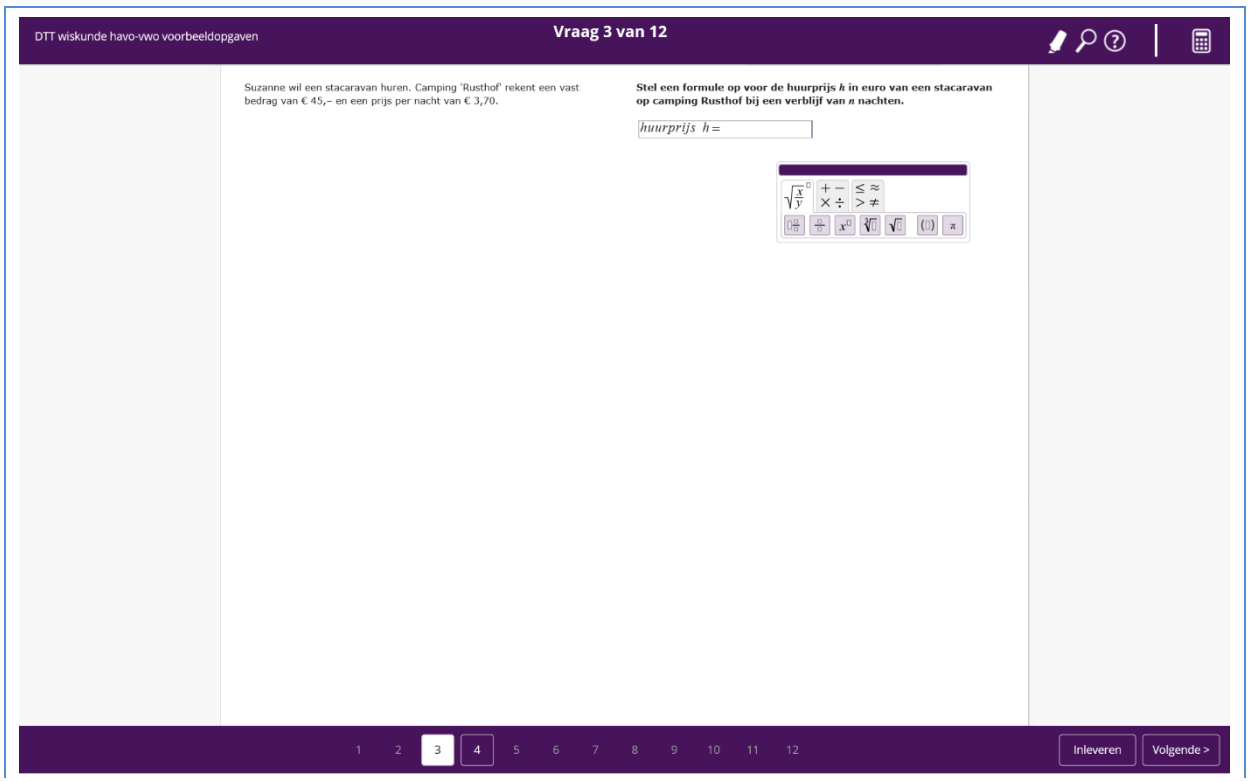


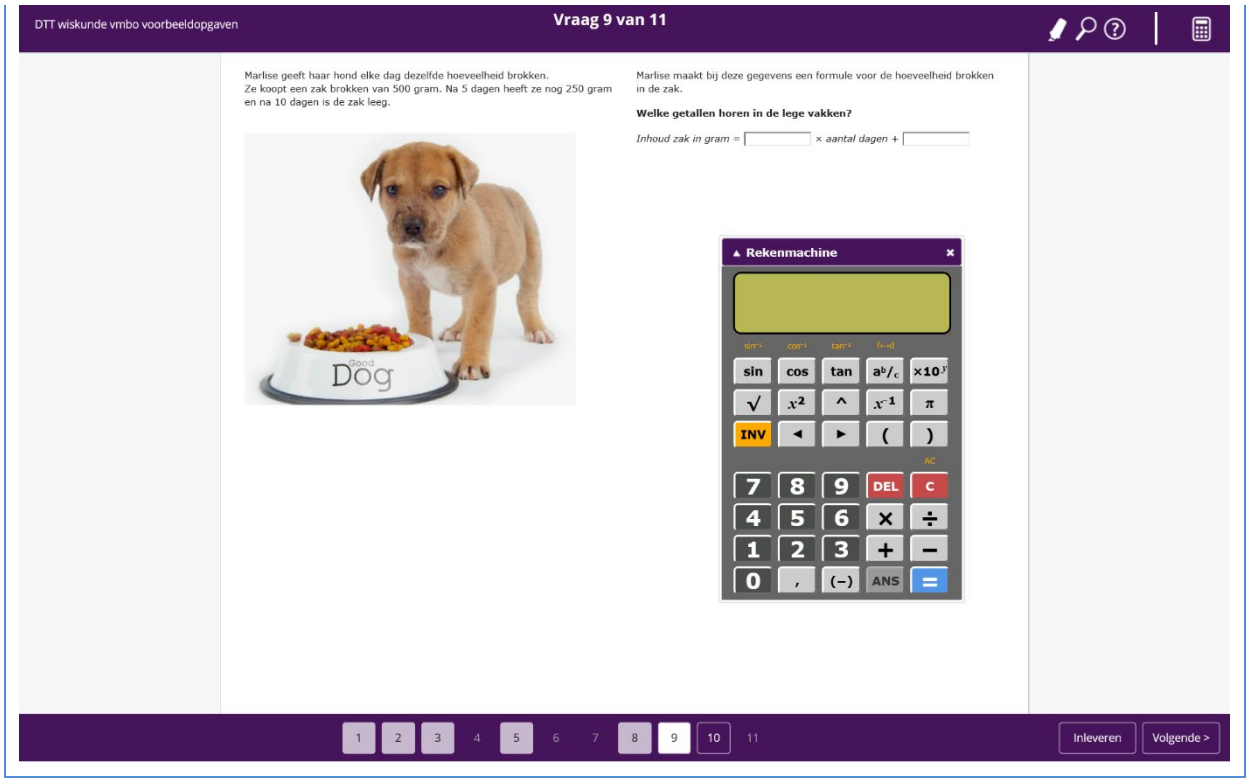
Short open-ended task

The student types in an answer. This type of task is used for languages, in order to test spelling. For mathematics tasks, students can use a special formula editor to enter input in the form of numbers and formulas.

Kind of interaction: Entering



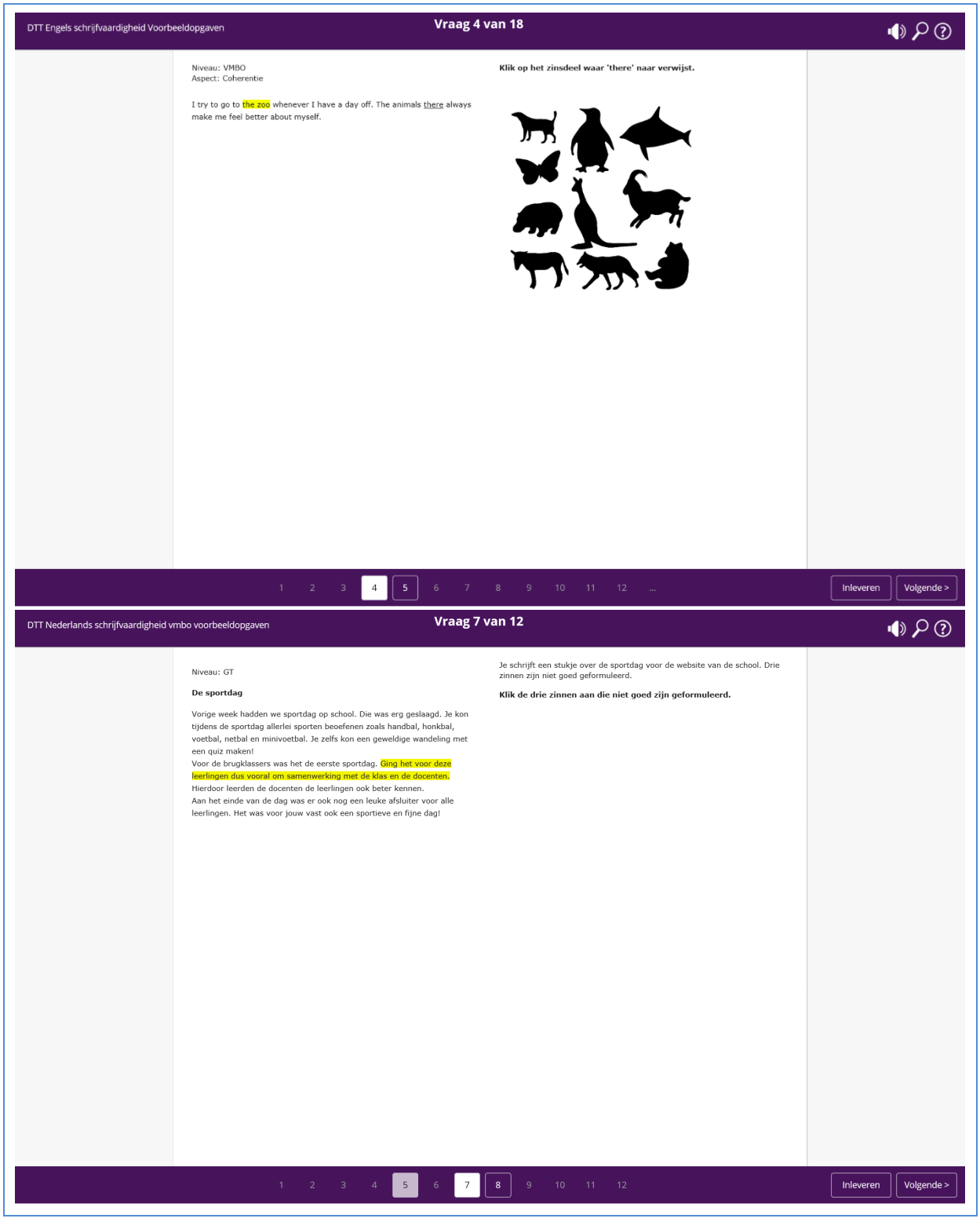



DTT wiskunde vmbo voorbeeldopgaven
Vraag 9 van 11
Marlise geeft haar hond elke dag dezelfde hoeveelheid brokken.
Ze koopt een zak brokken van 500 gram. Na 5 dagen heeft ze nog 250 gram en na 10 dagen is de zak leeg.
Good Dog
Marlise maakt bij deze gegevens een formule voor de hoeveelheid brokken in de zak.
Welke getallen horen in de lege vakken?
Inhoud zak in gram = × aantal dagen +
Rekenmachine
sin cos tan
INV
DEL
C
ANS
1 2 3 4 5 6 7 8 9 10 11
Inleveren
Volgende >

Marking task

The student selects active parts of a text. Suitable for finding answers in a text (e.g., a quotation or misspelled word).

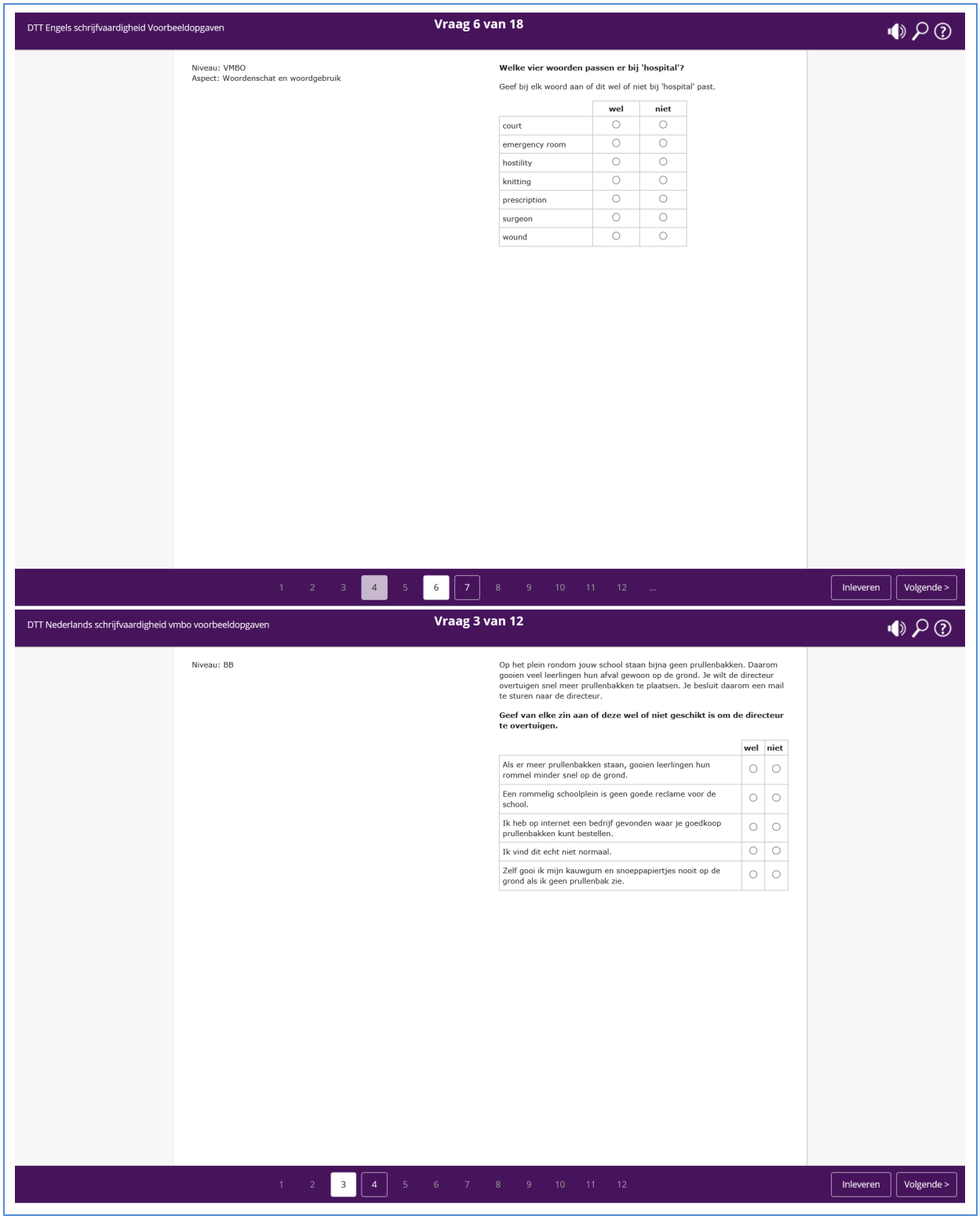
Kind of interaction: Clicking

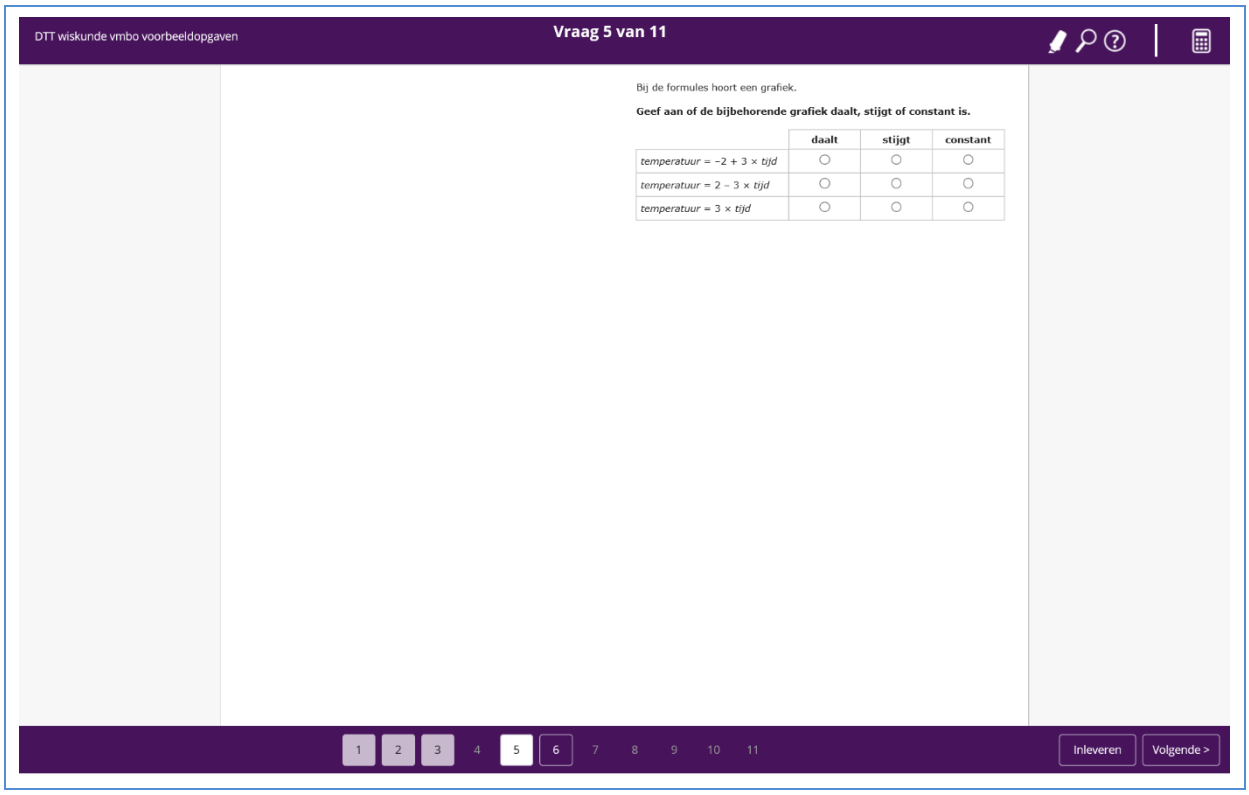


Matrix task

The student selects with several questions in a row one of (usually) two options (e.g., true or false). Suitable for asking about several details.

Kind of interaction: Clicking, multiple choice

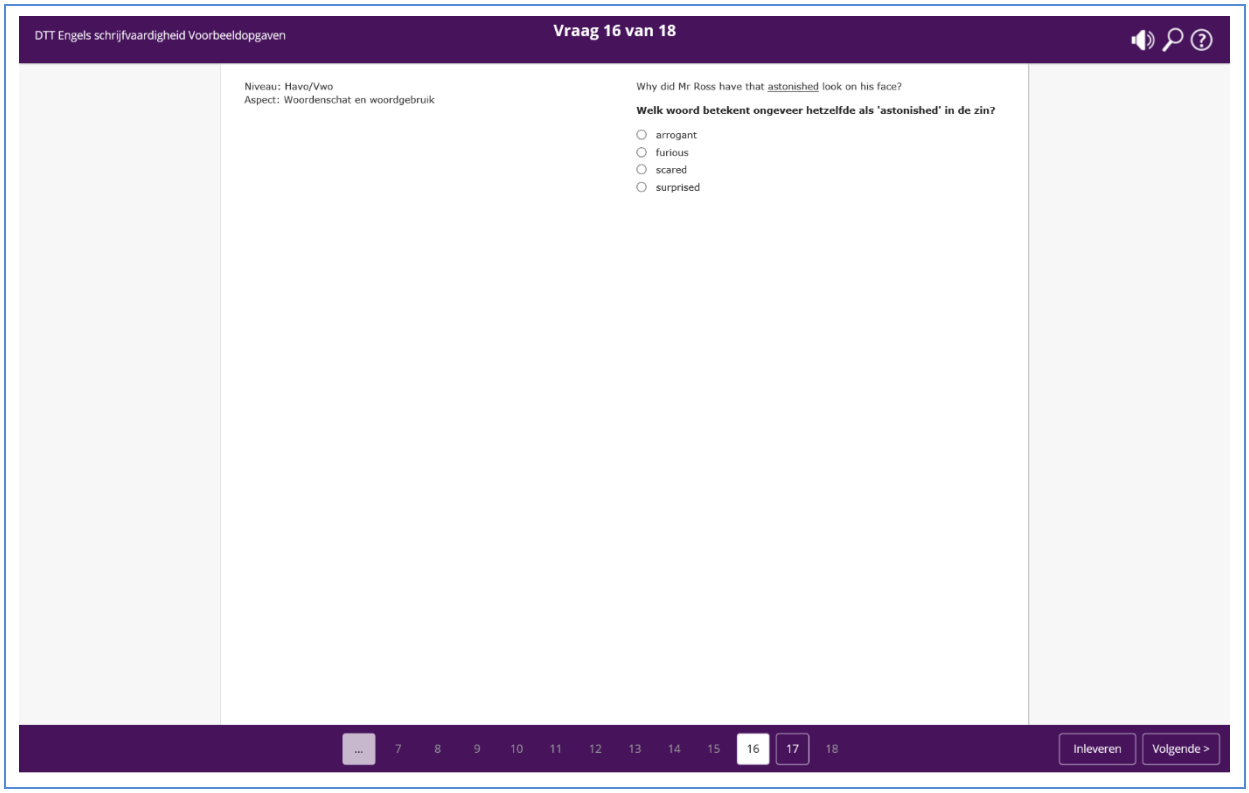


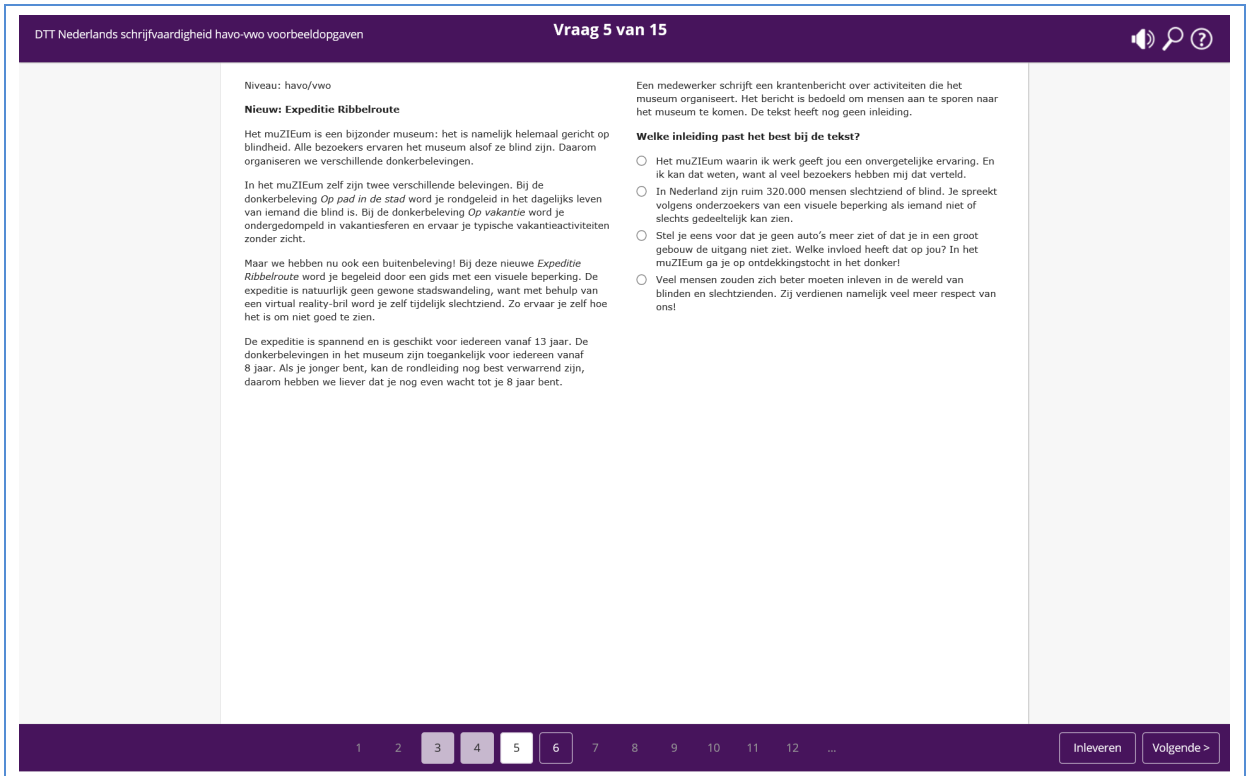

DTT wiskunde vmbo voorbeeldopgaven
Vraag 5 van 11
Bij de formules hoort een grafiek.
Geef aan of de bijbehorende grafiek daalt, stijgt of constant is.
daalt
stijgt
constant
temperatuur = −2 + 3 × tijd
temperatuur = 2 − 3 × tijd
temperatuur = 3 × tijd
1 2 3 4 5 6 7 8 9 10 11
Inleveren
Volgende >

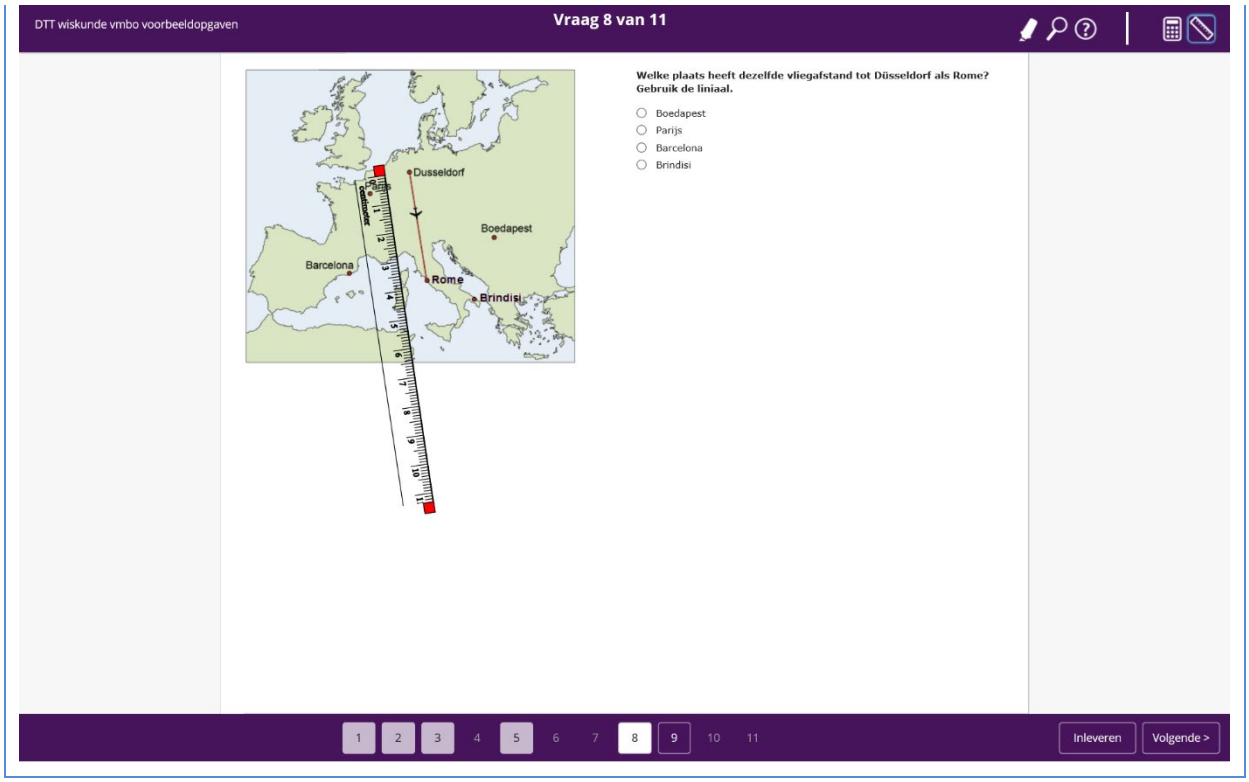
Multiple-choice task

The student selects the correct answer from a list of alternatives (texts, images, videos, audio fragments) or from active areas in an image.

Kind of interaction: Clicking, multiple choice



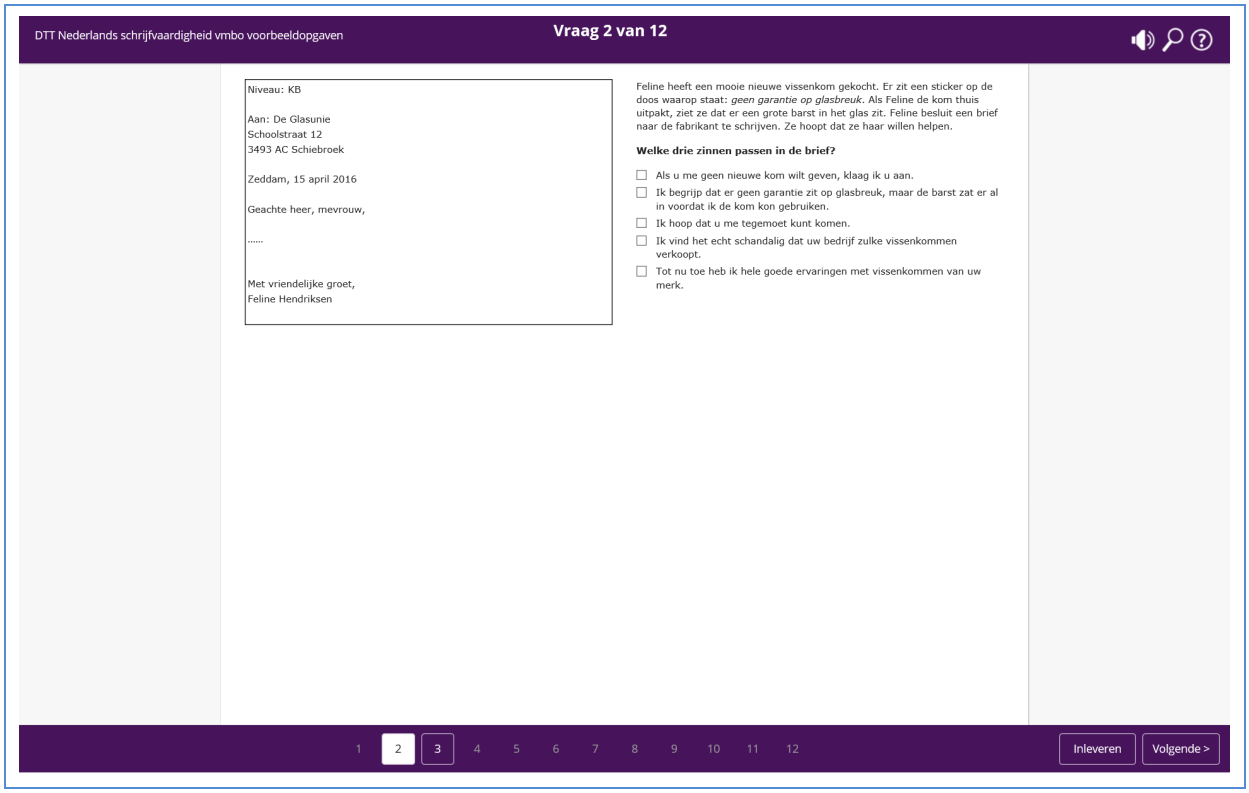



DTT wiskunde vmbo voorbeeldopgaven
Vraag 8 van 11
Welke plaats heeft dezelfde vliegafstand tot Düsseldorf als Rome? Gebruik de liniaal.
Boedapest
Parijs
Barcelona
Brindisi
Dusseldorf
Boedapest
Barcelona
Rome
Brindisi
1 2 3 4 5 6 7 8 9 10 11
Inleveren
Volgende >

Multiple-response task

The student selects one or more answers from a list of alternatives (texts, images, videos, audio fragments). This form is often used for tasks involving internet forms (then it is actually a kind of sorting question).

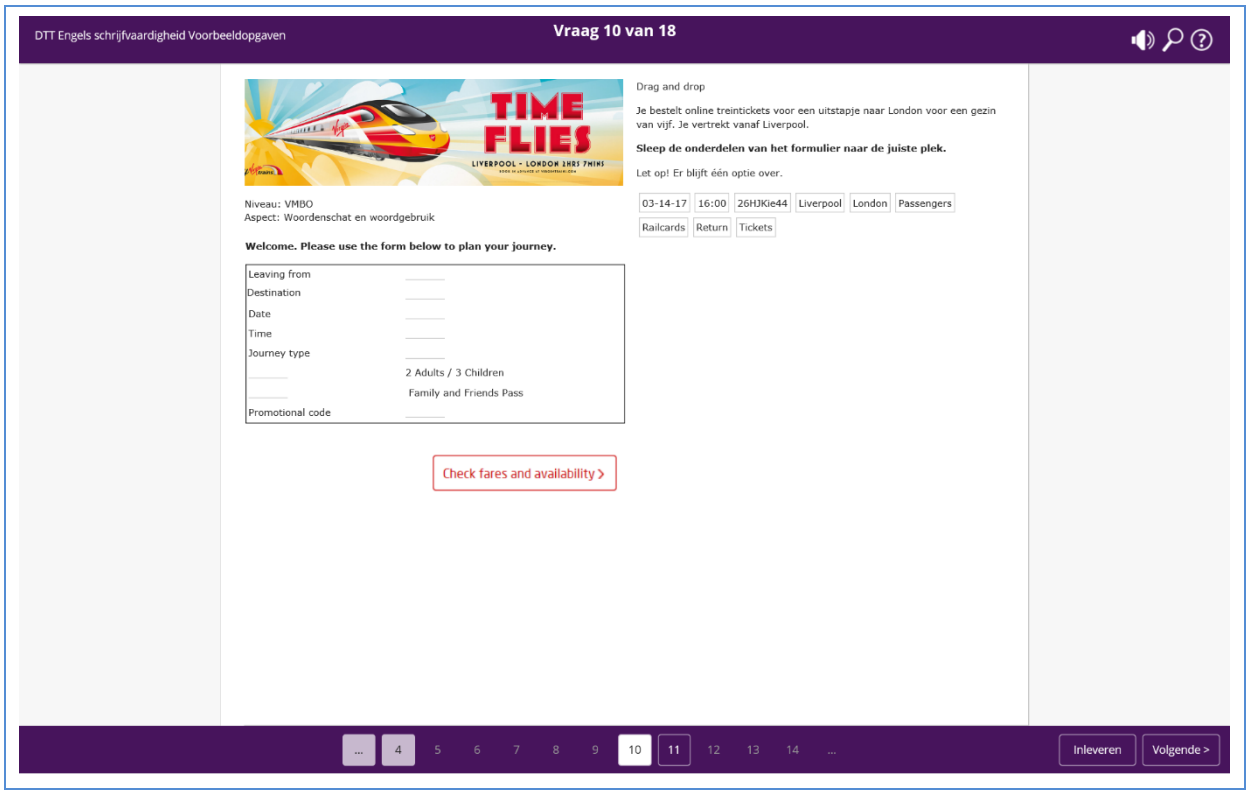
Kind of interaction: Clicking, multiple choice

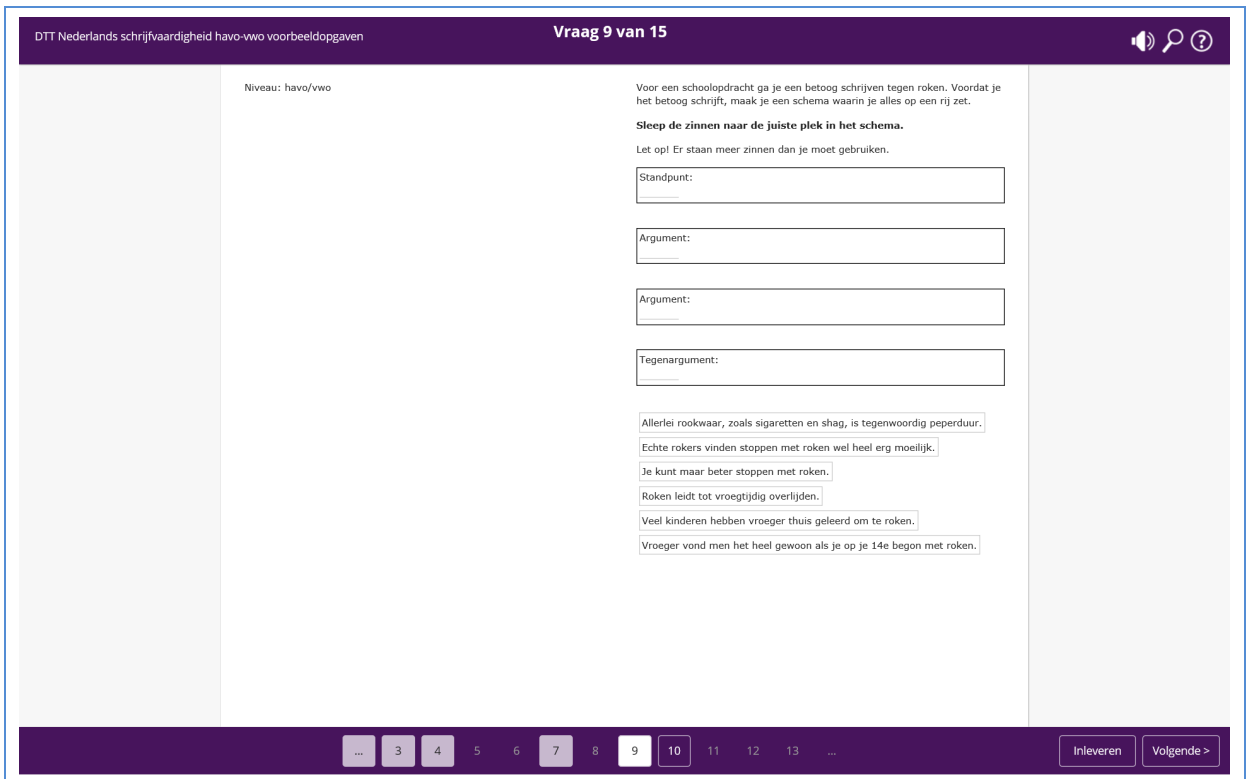


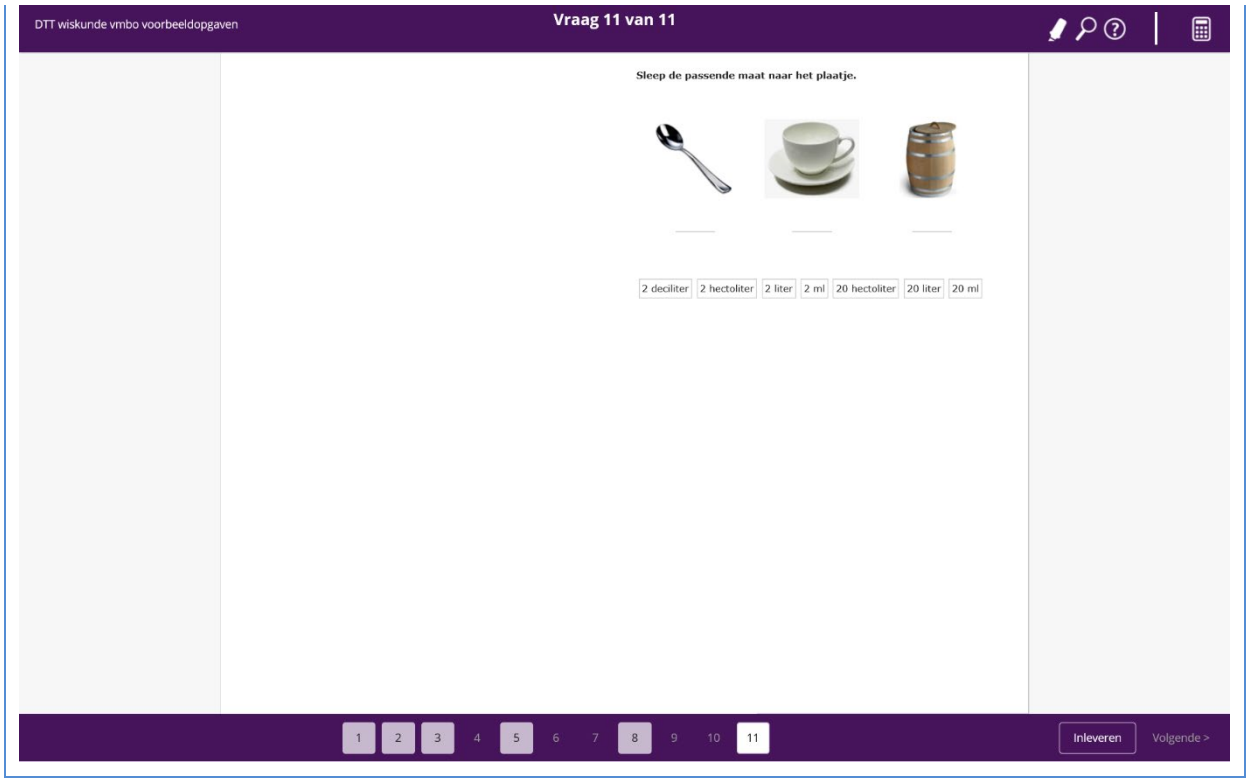
Dragging task

The student drags an answer in the form of text. Suitable for matching or sorting/categorizing.

Kind of interaction: Dragging



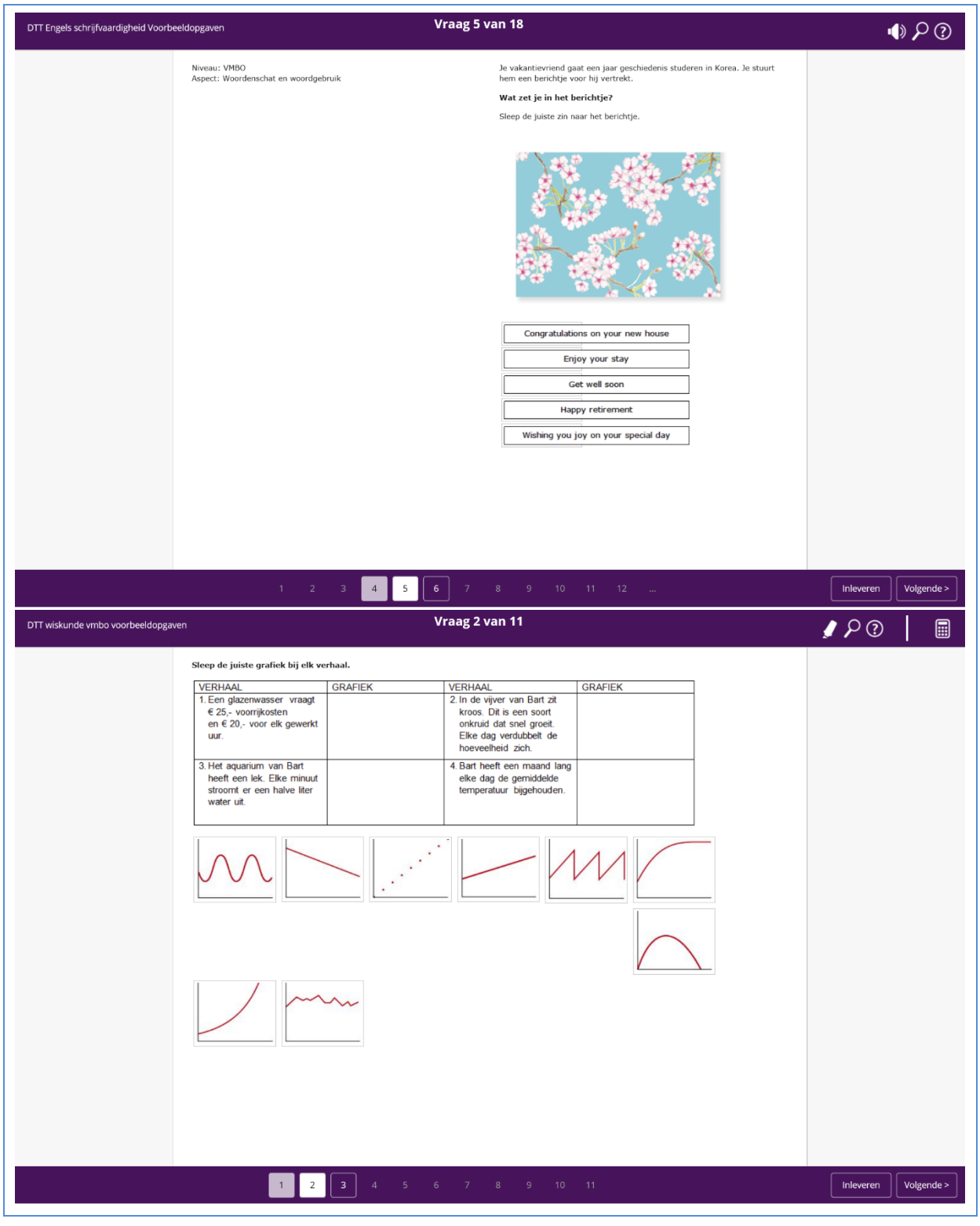



DTT wiskunde vmbo voorbeeldopgaven
Vraag 11 van 11
Sleep de passende maat naar het plaatje.
2 deciliter
2 hectoliter
2 liter
2 ml
20 hectoliter
20 liter
20 ml
1
2
3
4
5
6
7
8
9
10
11
Inleveren
Volgende >

Drag task with image

The student drags an answer. Suitable for linking images to words.

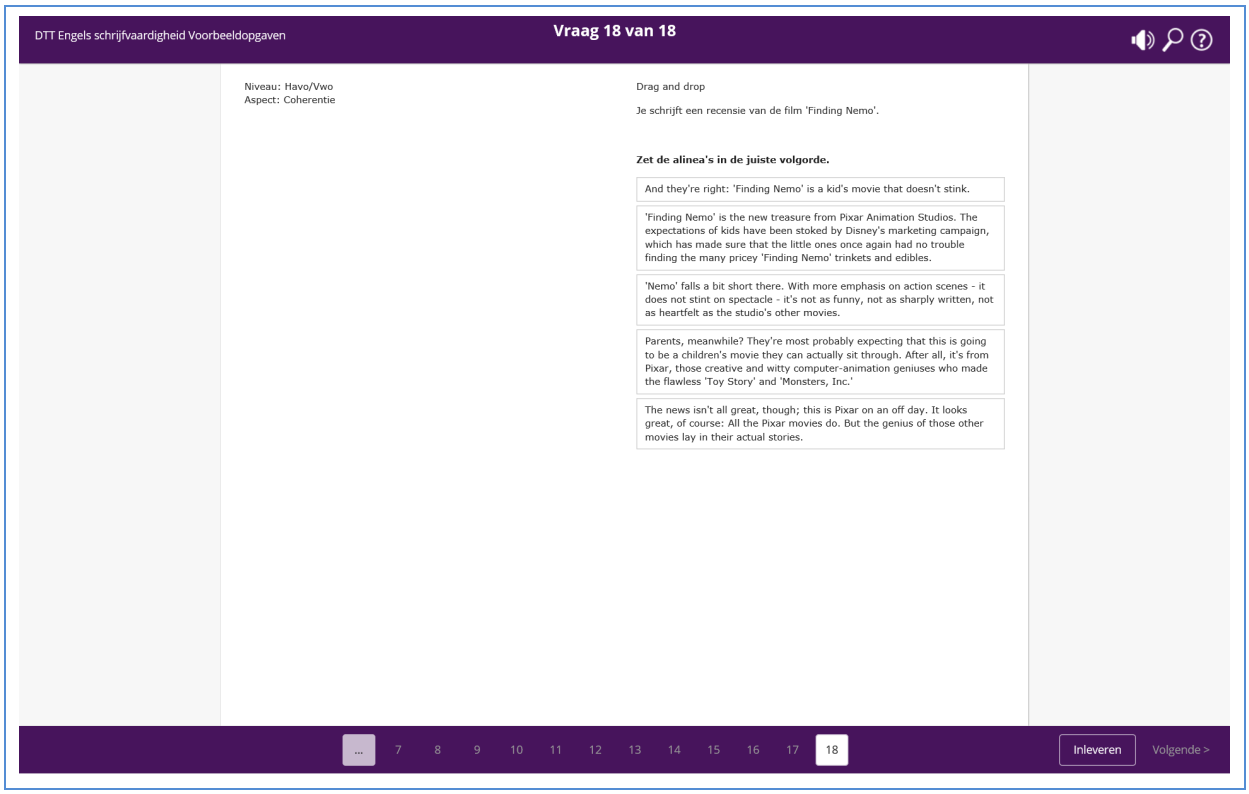
Kind of interaction: Dragging

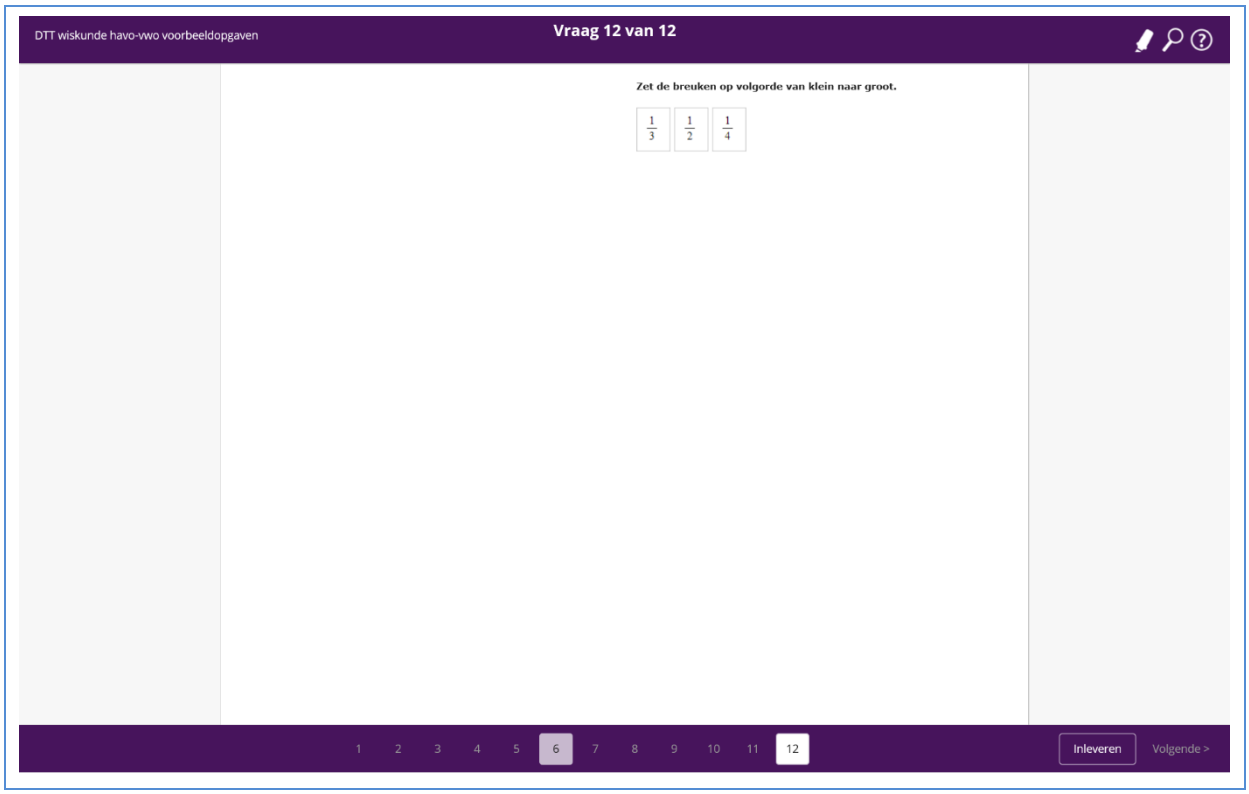


Ordering task

The student drags text fragments or numbers in a new order.

Kind of interaction: Dragging





17.2 Appendix Accuracy, boundary values, and prior-model probabilities for the block arrangements delivered

17.2.1 Results of the block arrangement for writing Dutch

Table 17-1. Percentages correct diagnoses writing Dutch, number of tasks and responses in the simulations for havo/vwo (excluding the tasks in the further-testing blocks)

Educational stream	Main aspects						Sub-aspects				
		Below level	At level	Above level	Number of tasks	Number of responses		Below level	At level	Above level	Number of responses
havo	M1	0.816	0.673	0.797	42	126	S11	0.767	0.604	0.637	64
							S12	0.616	0.624	0.764	37
							S13	0.816	0.673	0.797	25
	M2	0.805	0.767	0.727	67	160	S21	0.595	0.605	0.593	29
							S22	0.855	0.738	0.705	35
							S23	0.660	0.656	0.751	32
							S24	0.805	0.767	0.727	64
	M3	0.921	0.928	0.933	43	174	S31	0.799	0.653	0.603	60
							S32	0.682	0.611	0.634	39
							S33	0.853	0.841	0.927	75
	M4	0.950	0.950	0.950	36	204	S41	0.799	0.837	0.883	82
							S42	0.834	0.731	0.795	77
							S43	0.777	0.713	0.774	45
vwo	M1	0.964	0.829	0.719	50	106	S11	0.590	0.571	0.510	42
							S12	0.621	0.623	0.691	25
							S13	0.964	0.829	0.719	39
	M2	0.766	0.744	0.714	67	164	S21	0.672	0.687	0.667	43
							S22	0.879	0.775	0.710	56
							S23	0.766	0.744	0.714	40
							S24	0.564	0.596	0.583	25
	M3	0.916	0.916	0.916	51	168	S31	0.772	0.742	0.726	79
							S32	0.728	0.757	0.817	40
							S33	0.720	0.714	0.707	49
	M4	0.978	0.976	0.976	34	176	S41	0.946	0.877	0.910	82
							S42	0.778	0.739	0.834	59
							S43	0.835	0.806	0.875	35

Table 17-2. Percentages correct diagnoses writing Dutch, number of tasks and responses in the simulations for vmbo (excluding the tasks in the further-testing blocks)

Educational stream	Main aspects						Sub-aspects				
		Below level	At level	Above level	Number of tasks	Number of responses		Below level	At level	Above level	Number of responses
vmbo-bb	M1	0.950	0.949	0.949	43	160	S11	0.809	0.782	0.824	60
							S12	0.818	0.721	0.721	42
							S13	0.767	0.826	0.866	58
	M2	0.933	0.873	0.848	51	129	S21	0.844	0.720	0.660	44
							S22	0.615	0.604	0.666	19
							S23	0.933	0.873	0.848	66
	M3	0.955	0.955	0.954	47	157	S31	0.785	0.775	0.847	29
							S32	0.736	0.834	0.855	84
							S33	0.865	0.769	0.759	44
	M4	0.949	0.944	0.944	40	213	S41	0.878	0.712	0.702	65
							S42	0.738	0.774	0.831	91
							S43	0.921	0.851	0.863	57
vmbo-kb	M1	0.936	0.935	0.935	44	190	S11	0.726	0.797	0.853	59
							S12	0.861	0.712	0.720	44
							S13	0.853	0.783	0.814	87
	M2	0.906	0.906	0.906	56	118	S21	0.651	0.662	0.732	28
							S22	0.607	0.676	0.627	21
							S23	0.884	0.804	0.851	69
	M3	0.970	0.968	0.968	50	192	S31	0.725	0.811	0.898	62
							S32	0.830	0.803	0.882	74
							S33	0.887	0.748	0.800	56
	M4	0.962	0.957	0.957	37	212	S41	0.766	0.722	0.734	65
							S42	0.856	0.836	0.867	101
							S43	0.905	0.822	0.869	46
vmbo-gt	M1	0.890	0.890	0.890	43	170	S11	0.807	0.737	0.772	56
							S12	0.716	0.699	0.656	59
							S13	0.818	0.761	0.759	55
	M2	0.952	0.945	0.947	64	127	S21	0.747	0.660	0.772	27
							S22	0.602	0.609	0.699	24
							S23	0.966	0.883	0.885	76
	M3	0.947	0.945	0.945	43	163	S31	0.827	0.781	0.815	57
							S32	0.738	0.808	0.857	61
							S33	0.846	0.705	0.786	45
	M4	0.859	0.789	0.736	41	212	S41	0.710	0.701	0.717	66
							S42	0.777	0.790	0.861	92
							S43	0.859	0.789	0.736	54

Table 17-3. Exit probabilities writing Dutch for havo/vwo

Educational stream	Main aspects				Sub-aspects			
		Below level	At level	Above level		Below level	At level	Above level
havo	M1	0.010	0.947	0.766	S11	0.605	0.953	0.990
					S12	0.644	0.990	0.867
					S13	0.567	0.990	0.912
	M2	0.010	0.841	0.010	S21	0.796	0.877	0.920
					S22	0.640	0.953	0.902
					S23	0.727	0.990	0.802
					S24	0.666	0.948	0.884
	M3	0.620	0.864	0.010	S31	0.922	0.971	0.590
					S32	0.944	0.980	0.010
					S33	0.860	0.922	0.010
	M4	0.854	0.956	0.737	S41	0.318	0.405	0.277
					S42	0.892	0.694	0.990
					S43	0.878	0.990	0.889
vwo	M1	0.010	0.921	0.778	S11	0.010	0.947	0.990
					S12	0.847	0.947	0.918
					S13	0.010	0.973	0.990
	M2	0.010	0.882	0.691	S21	0.490	0.978	0.885
					S22	0.010	0.967	0.980
					S23	0.010	0.968	0.902
					S24	0.620	0.942	0.959
	M3	0.741	0.922	0.827	S31	0.279	0.419	0.302
					S32	0.990	0.634	0.851
					S33	0.836	0.667	0.990
	M4	0.010	0.955	0.694	S41	0.240	0.416	0.344
					S42	0.731	0.990	0.903
					S43	0.647	0.960	0.831

Table 17-4. Exit probabilities writing Dutch for vmbo

Educational stream	Main aspects				Sub-aspects			
		Below level	At level	Above level		Below level	At level	Above level
vmbo-bb	M1	0.010	0.978	0.872	S11	0.207	0.434	0.360
					S12	0.990	0.990	0.990
					S13	0.620	0.846	0.760
	M2	0.010	0.840	0.590	S21	0.010	0.971	0.990
					S22	0.990	0.913	0.903
					S23	0.010	0.968	0.787
	M3	0.010	0.960	0.657	S31	0.228	0.425	0.347
					S32	0.726	0.968	0.908
					S33	0.413	0.895	0.795
	M4	0.010	0.997	0.784	S41	0.170	0.430	0.400
					S42	0.990	0.990	0.917
					S43	0.010	0.994	0.864
vmbo-kb	M1	0.010	0.994	0.874	S11	0.176	0.442	0.382
					S12	0.010	0.990	0.844
					S13	0.010	0.971	0.990
	M2	0.010	0.982	0.909	S21	0.200	0.427	0.373
					S22	0.990	0.990	0.784
					S23	0.010	0.982	0.990
	M3	0.587	0.977	0.762	S31	0.226	0.402	0.372
					S32	0.746	0.905	0.878
					S33	0.990	0.618	0.571
	M4	0.010	0.984	0.809	S41	0.181	0.421	0.398
					S42	0.010	0.934	0.929
					S43	0.010	0.989	0.871
vmbo-gt	M1	0.010	0.991	0.910	S11	0.172	0.453	0.375
					S12	0.990	0.990	0.990
					S13	0.010	0.990	0.990
	M2	0.010	0.765	0.506	S21	0.010	0.953	0.990
					S22	0.010	0.893	0.851
					S23	0.010	0.913	0.801
	M3	0.010	0.983	0.869	S31	0.191	0.417	0.392
					S32	0.508	0.752	0.757
					S33	0.010	0.948	0.932
	M4	0.010	0.795	0.544	S41	0.552	0.986	0.914
					S42	0.010	0.990	0.809
					S43	0.010	0.935	0.985

Table 17-5. Prior model probabilities writing Dutch

Educational stream		Main aspects			Sub-aspects		
		Below level	At level	Above level	Below level	At level	Above level
havo	M1	0.136	0.482	0.382	0.280	0.380	0.330
	M2	0.172	0.498	0.330	0.340	0.400	0.260
	M3	0.511	0.449	0.040	0.422	0.420	0.220
	M4	0.302	0.477	0.221	0.318	0.405	0.277
vwo	M1	0.046	0.552	0.402	0.400	0.443	0.230
	M2	0.069	0.512	0.419	0.290	0.430	0.270
	M3	0.224	0.505	0.271	0.279	0.419	0.302
	M4	0.147	0.498	0.355	0.240	0.416	0.344
vmbo-bb	M1	0.080	0.534	0.386	0.207	0.434	0.360
	M2	0.066	0.521	0.413	0.300	0.420	0.280
	M3	0.123	0.516	0.361	0.228	0.425	0.347
	M4	0.007	0.527	0.466	0.170	0.430	0.400
vmbo-kb	M1	0.018	0.551	0.430	0.176	0.442	0.382
	M2	0.067	0.520	0.414	0.200	0.427	0.373
	M3	0.119	0.471	0.410	0.226	0.402	0.372
	M4	0.029	0.508	0.463	0.181	0.421	0.398
vmbo-gt	M1	0.011	0.572	0.417	0.172	0.453	0.375
	M2	0.054	0.547	0.399	0.330	0.500	0.366
	M3	0.049	0.500	0.451	0.191	0.417	0.392
	M4	0.084	0.558	0.358	0.250	0.420	0.330

17.2.2 Results of the block arrangement for writing English

Table 17-6. Percentages correct diagnoses writing English, number of tasks and responses in the simulations for havo/vwo (excluding the tasks in the further-testing blocks)

Educational stream	Main aspects						Sub-aspects				
		Below level	At level	Above level	Number of tasks	Number of responses		Below level	At level	Above level	Number of responses
havo	M1	0.932	0.932	0.932	48	149	S11	0.822	0.753	0.787	55
							S12	0.841	0.852	0.907	94
	M2	0.913	0.913	0.913	54	84	S21	0.827	0.780	0.864	42
							S22	0.841	0.761	0.802	42
	M3	0.946	0.946	0.947	78	130	S31	0.896	0.863	0.883	75
							S32	0.821	0.811	0.851	55
	M4	0.933	0.933	0.933	73	109	S41	0.880	0.844	0.877	58
							S42	0.822	0.777	0.819	51
vwo	M1	0.882	0.882	0.882	39	100	S11	0.627	0.641	0.629	54
							S12	0.855	0.830	0.870	46
	M2	0.925	0.925	0.925	51	92	S21	0.812	0.728	0.745	37
							S22	0.847	0.849	0.909	55
	M3	0.917	0.917	0.917	72	114	S31	0.812	0.765	0.800	57
							S32	0.827	0.808	0.872	57
	M4	0.876	0.876	0.876	52	82	S41	0.835	0.783	0.836	44
							S42	0.738	0.657	0.707	38

Table 17-7. Percentages correct diagnoses writing English, number of tasks and responses in the simulations for vmbo (excluding the tasks in the further-testing blocks)

Educational stream	Main aspects						Sub-aspects				
		Below level	At level	Above level	Number of tasks	Number of responses		Below level	At level	Above level	Number of responses
vmbo-bb	M1	0.903	0.903	0.903	51	107	S11	0.744	0.710	0.725	54
							S12	0.863	0.820	0.861	53
	M2	0.941	0.941	0.941	61	80	S21	0.873	0.867	0.891	42
							S22	0.837	0.799	0.809	38
	M3	0.964	0.965	0.966	58	129	S31	0.923	0.906	0.945	80
							S32	0.813	0.838	0.864	49
	M4	0.929	0.928	0.928	65	70	S41	0.881	0.828	0.874	33
							S42	0.818	0.771	0.790	37
vmbo-kb	M1	0.890	0.890	0.890	52	124	S11	0.796	0.756	0.835	64
							S12	0.811	0.747	0.777	60
	M2	0.954	0.954	0.954	65	99	S21	0.842	0.848	0.891	45
							S22	0.898	0.852	0.883	54
	M3	0.971	0.971	0.971	68	147	S31	0.942	0.897	0.929	82
							S32	0.839	0.841	0.911	65
	M4	0.960	0.960	0.960	70	109	S41	0.828	0.863	0.944	60
							S42	0.914	0.843	0.824	49
vmbo-gt	M1	0.956	0.956	0.956	43	134	S11	0.906	0.861	0.868	88
							S12	0.840	0.829	0.904	46
	M2	0.961	0.961	0.962	73	96	S21	0.888	0.851	0.873	40
							S22	0.890	0.874	0.912	56
	M3	0.960	0.959	0.960	60	131	S31	0.927	0.890	0.922	81
							S32	0.806	0.807	0.879	50
	M4	0.936	0.936	0.936	70	99	S41	0.845	0.809	0.863	48
							S42	0.850	0.838	0.866	51

Table 17-8. Exit probabilities writing English for havo/vwo

Educational stream	Main aspects				Sub-aspects			
		Below level	At level	Above level		Below level	At level	Above level
havo	M1	0.599	0.972	0.813	S11	0.691	0.802	0.811
					S12	0.792	0.920	0.911
	M2	0.714	0.957	0.863	S21	0.823	0.945	0.718
					S22	0.827	0.990	0.698
	M3	0.763	0.948	0.811	S31	0.855	0.855	0.860
					S32	0.875	0.770	0.615
	M4	0.813	0.961	0.591	S41	0.868	0.922	0.831
					S42	0.807	0.897	0.813
vwo	M1	0.748	0.990	0.919	S11	0.752	0.990	0.708
					S12	0.990	0.605	0.587
	M2	0.825	0.976	0.757	S21	0.867	0.753	0.649
					S22	0.880	0.881	0.783
	M3	0.767	0.961	0.874	S31	0.921	0.990	0.607
					S32	0.850	0.844	0.829
	M4	0.743	0.876	0.894	S41	0.750	0.781	0.906
					S42	0.792	0.990	0.850

Table 17-9. Exit probabilities writing English for vmbo

Educational stream	Main aspects				Sub-aspects			
		Below level	At level	Above level		Below level	At level	Above level
vmbo-bb	M1	0.761	0.937	0.812	S11	0.620	0.990	0.831
					S12	0.685	0.803	0.848
	M2	0.736	0.970	0.796	S21	0.756	0.929	0.888
					S22	0.796	0.684	0.990
	M3	0.694	0.957	0.010	S31	0.770	0.967	0.564
					S32	0.872	0.990	0.481
	M4	0.774	0.894	0.742	S41	0.807	0.814	0.820
					S42	0.557	0.990	0.880
vmbo-kb	M1	0.806	0.882	0.864	S11	0.813	0.990	0.901
					S12	0.990	0.990	0.906
	M2	0.726	0.966	0.778	S21	0.850	0.854	0.831
					S22	0.803	0.990	0.764
	M3	0.010	0.922	0.651	S31	0.010	0.936	0.844
					S32	0.895	0.876	0.862
	M4	0.698	0.958	0.646	S41	0.896	0.990	0.519
					S42	0.753	0.954	0.692
vmbo-gt	M1	0.721	0.965	0.617	S11	0.878	0.990	0.990
					S12	0.795	0.922	0.659
	M2	0.792	0.965	0.666	S21	0.896	0.755	0.990
					S22	0.885	0.990	0.782
	M3	0.743	0.963	0.739	S31	0.794	0.922	0.753
					S32	0.854	0.907	0.839
	M4	0.784	0.966	0.717	S41	0.923	0.888	0.719
					S42	0.867	0.858	0.808

Table 17-10. Prior model probabilities writing English

Educational stream	Main aspects	Sub-aspects					
		Below level	At level	Above level	Below level	At level	Above level
havo	M1	0.171	0.479	0.350	0.330	0.340	0.330
	M2	0.232	0.507	0.262	0.330	0.340	0.330
	M3	0.286	0.504	0.209	0.330	0.340	0.330
	M4	0.288	0.502	0.210	0.330	0.340	0.330
vwo	M1	0.193	0.522	0.285	0.330	0.340	0.330
	M2	0.303	0.521	0.176	0.330	0.340	0.330
	M3	0.300	0.493	0.207	0.330	0.340	0.330
	M4	0.218	0.451	0.330	0.330	0.340	0.330
vmbo-bb	M1	0.263	0.509	0.228	0.330	0.340	0.330
	M2	0.219	0.525	0.256	0.330	0.340	0.330
	M3	0.379	0.506	0.115	0.330	0.340	0.330
	M4	0.234	0.495	0.271	0.330	0.340	0.330
vmbo-kb	M1	0.174	0.493	0.333	0.330	0.340	0.330
	M2	0.253	0.507	0.239	0.330	0.340	0.330
	M3	0.260	0.486	0.254	0.330	0.340	0.330
	M4	0.352	0.504	0.144	0.330	0.340	0.330
vmbo-gt	M1	0.326	0.479	0.195	0.330	0.340	0.330
	M2	0.351	0.479	0.170	0.330	0.340	0.330
	M3	0.273	0.499	0.228	0.330	0.340	0.330
	M4	0.324	0.508	0.168	0.330	0.340	0.330

17.2.4 Results of the block arrangement for mathematics

Table 17-11. Percentages correct diagnoses Mathematics, number of tasks and responses in the simulations for havo/vwo (excluding the tasks in the further-testing blocks)

Educational stream	Main aspects						Sub-aspects				
		Below level	At level	Above level	Number of tasks	Number of responses		Below level	At level	Above level	Number of responses
havo	B	0.790	0.790	0.790	34	69	B1	0.667	0.526	0.712	28
							B2	0.513	0.603	0.735	20
							B3	0.670	0.612	0.878	21
	C	0.800	0.800	0.800	38	47	C1	0.647	0.612	0.720	11
							C2	0.667	0.371	0.749	15
							C3	0.641	0.582	0.739	21
	D	0.774	0.773	0.774	37	49	D1	0.730	0.399	0.689	18
							D2	0.751	0.344	0.635	14
							D3	0.751	0.462	0.726	17
	E	0.820	0.820	0.820	36	56	E1	0.828	0.571	0.858	15
							E2	0.692	0.397	0.745	13
							E3	0.674	0.436	0.795	28
	F	0.767	0.767	0.767	41	56	F1	0.778	0.525	0.736	19
							F2	0.594	0.533	0.629	23
							F3	0.580	0.360	0.599	14
vwo	B	0.779	0.779	0.779	38	67	B1	0.637	0.362	0.676	20
							B2	0.396	0.572	0.667	18
							B3	0.728	0.571	0.867	29
	C	0.774	0.774	0.774	34	46	C1	0.558	0.485	0.722	15
							C2	0.688	0.429	0.769	11
							C3	0.692	0.423	0.777	20
	D	0.847	0.847	0.848	41	55	D1	0.777	0.514	0.840	18
							D2	0.733	0.477	0.701	17
							D3	0.749	0.541	0.810	20
	E	0.853	0.852	0.853	42	48	E1	0.803	0.664	0.839	18
							E2	0.770	0.461	0.715	17
							E3	0.722	0.459	0.690	13
	F	0.754	0.754	0.754	36	48	F1	0.649	0.417	0.612	16
							F2	0.626	0.657	0.791	23
							F3	0.445	0.317	0.667	9

Table 17-12. Percentages correct diagnoses Mathematics, number of tasks and responses in the simulations for vmbo (excluding the tasks in the further-testing blocks)

Educational stream	Main aspects						Sub-aspects				
		Below level	At level	Above level	Number of tasks	Number of responses		Below level	At level	Above level	Number of responses
vmbo-bb	B	0.878	0.878	0.878	70	90	B1	0.750	0.576	0.716	37
							B2	0.742	0.566	0.794	31
							B3	0.814	0.605	0.852	22
	C	0.862	0.862	0.862	56	76	C1	0.859	0.721	0.833	32
							C2	0.606	0.523	0.619	25
							C3	0.658	0.380	0.678	19
	D	0.835	0.835	0.836	49	77	D1	0.733	0.482	0.761	23
							D2	0.791	0.618	0.846	42
							D3	0.735	0.355	0.553	12
	E	0.795	0.795	0.795	43	49	E1	0.736	0.447	0.703	20
							E2	0.648	0.384	0.585	10
							E3	0.746	0.567	0.800	19
vmbo-kb	B	0.805	0.805	0.805	57	84	B1	0.716	0.508	0.691	28
							B2	0.673	0.564	0.715	36
							B3	0.657	0.502	0.714	20
	C	0.814	0.814	0.814	46	70	C1	0.849	0.643	0.755	27
							C2	0.672	0.368	0.647	24
							C3	0.656	0.337	0.595	19
	D	0.836	0.836	0.836	55	70	D1	0.674	0.446	0.763	26
							D2	0.770	0.582	0.804	25
							D3	0.693	0.516	0.725	19
	E	0.886	0.886	0.887	41	63	E1	0.817	0.589	0.711	26
							E2	0.741	0.380	0.752	17
							E3	0.795	0.635	0.887	20
vmbo-gt	B	0.826	0.826	0.826	53	70	B1	0.719	0.585	0.801	33
							B2	0.679	0.603	0.673	18
							B3	0.745	0.477	0.814	19
	C	0.833	0.833	0.833	67	77	C1	0.807	0.645	0.855	29
							C2	0.604	0.406	0.726	25
							C3	0.699	0.482	0.774	23
	D	0.838	0.838	0.838	52	63	D1	0.767	0.561	0.819	23
							D2	0.756	0.457	0.789	21
							D3	0.743	0.493	0.764	19
	E	0.872	0.872	0.872	46	61	E1	0.799	0.657	0.846	28
							E2	0.775	0.470	0.705	15
							E3	0.797	0.457	0.772	18

Table 17-13. Exit probabilities Mathematics for havo/vwo

Educational stream	Main aspects				Sub-aspects			
		Below level	At level	Above level		Below level	At level	Above level
havo	B	0.939	0.877	0.010	B1	0.580	0.990	0.990
					B2	0.963	0.608	0.990
					B3	0.847	0.682	0.010
	C	0.872	0.734	0.846	C1	0.792	0.990	0.817
					C2	0.438	0.990	0.990
					C3	0.461	0.470	0.990
	D	0.674	0.990	0.873	D1	0.463	0.990	0.990
					D2	0.990	0.990	0.912
					D3	0.990	0.990	0.576
	E	0.990	0.857	0.697	E1	0.657	0.736	0.788
					E2	0.990	0.672	0.677
					E3	0.990	0.990	0.990
	F	0.894	0.990	0.894	F1	0.816	0.799	0.541
					F2	0.990	0.990	0.990
					F3	0.990	0.990	0.684
vwo	B	0.927	0.924	0.783	B1	0.603	0.990	0.462
					B2	0.474	0.990	0.990
					B3	0.921	0.990	0.579
	C	0.866	0.773	0.803	C1	0.474	0.990	0.398
					C2	0.576	0.990	0.445
					C3	0.731	0.990	0.990
	D	0.889	0.788	0.830	D1	0.884	0.671	0.546
					D2	0.990	0.990	0.406
					D3	0.771	0.990	0.990
	E	0.865	0.905	0.695	E1	0.833	0.990	0.828
					E2	0.471	0.990	0.610
					E3	0.990	0.990	0.990
	F	0.925	0.719	0.562	F1	0.643	0.990	0.990
					F2	0.923	0.990	0.990
					F3	0.792	0.990	0.990

Table 17-14. Exit probabilities Mathematics for vmbo

Educational stream	Main aspects				Sub-aspects			
		Below level	At level	Above level		Below level	At level	Above level
vmbo-bb	B	0.893	0.919	0.740	B1	0.990	0.608	0.990
					B2	0.990	0.717	0.578
					B3	0.857	0.875	0.824
	C	0.898	0.923	0.858	C1	0.931	0.990	0.869
					C2	0.990	0.990	0.431
					C3	0.990	0.990	0.990
	D	0.898	0.949	0.800	D1	0.870	0.648	0.990
					D2	0.990	0.911	0.798
					D3	0.990	0.990	0.421
	E	0.765	0.797	0.869	E1	0.463	0.570	0.601
					E2	0.470	0.990	0.477
					E3	0.518	0.608	0.643
vmbo-kb	B	0.908	0.895	0.866	B1	0.821	0.990	0.838
					B2	0.659	0.990	0.849
					B3	0.990	0.990	0.990
	C	0.861	0.790	0.909	C1	0.762	0.547	0.564
					C2	0.990	0.990	0.990
					C3	0.990	0.639	0.907
	D	0.912	0.814	0.809	D1	0.522	0.990	0.990
					D2	0.748	0.990	0.826
					D3	0.757	0.990	0.493
	E	0.905	0.925	0.847	E1	0.764	0.596	0.586
					E2	0.990	0.990	0.990
					E3	0.661	0.787	0.653
vmbo-gt	B	0.943	0.858	0.682	B1	0.630	0.990	0.822
					B2	0.990	0.990	0.481
					B3	0.990	0.621	0.687
	C	0.913	0.938	0.509	C1	0.752	0.857	0.523
					C2	0.990	0.990	0.990
					C3	0.737	0.674	0.726
	D	0.880	0.990	0.700	D1	0.701	0.990	0.747
					D2	0.584	0.649	0.990
					D3	0.634	0.990	0.990
	E	0.906	0.873	0.871	E1	0.472	0.990	0.826
					E2	0.713	0.990	0.990
					E3	0.530	0.990	0.990

Table 17-15. Prior model probabilities Mathematics

Educational stream	Main aspects				Sub-aspects		
		Below level	At level	Above level	Below level	At level	Above level
havo	B	0.495	0.472	0.034	0.330	0.340	0.330
	C	0.353	0.436	0.211	0.330	0.340	0.330
	D	0.306	0.435	0.259	0.330	0.340	0.330
	E	0.394	0.486	0.120	0.330	0.340	0.330
	F	0.336	0.444	0.220	0.330	0.340	0.330
vwo	B	0.454	0.426	0.120	0.330	0.340	0.330
	C	0.363	0.418	0.219	0.330	0.340	0.330
	D	0.351	0.463	0.186	0.330	0.340	0.330
	E	0.370	0.468	0.163	0.330	0.340	0.330
	F	0.430	0.405	0.165	0.330	0.340	0.330
vmbo-bb	B	0.370	0.471	0.159	0.330	0.340	0.330
	C	0.304	0.444	0.252	0.330	0.340	0.330
	D	0.321	0.461	0.218	0.330	0.340	0.330
	E	0.288	0.454	0.258	0.330	0.340	0.330
vmbo-kb	B	0.305	0.429	0.266	0.330	0.340	0.330
	C	0.170	0.457	0.373	0.330	0.340	0.330
	D	0.340	0.447	0.213	0.330	0.340	0.330
	E	0.318	0.472	0.210	0.330	0.340	0.330
vmbo-gt	B	0.382	0.477	0.141	0.330	0.340	0.330
	C	0.413	0.480	0.107	0.330	0.340	0.330
	D	0.342	0.504	0.154	0.330	0.340	0.330
	E	0.290	0.463	0.246	0.330	0.340	0.330

17.3 Appendix Visualizations of the paths followed, by subject and educational stream (for each package)

17.3.1 Visualizations of the paths followed in the adaptive assessment 2017 of writing Dutch

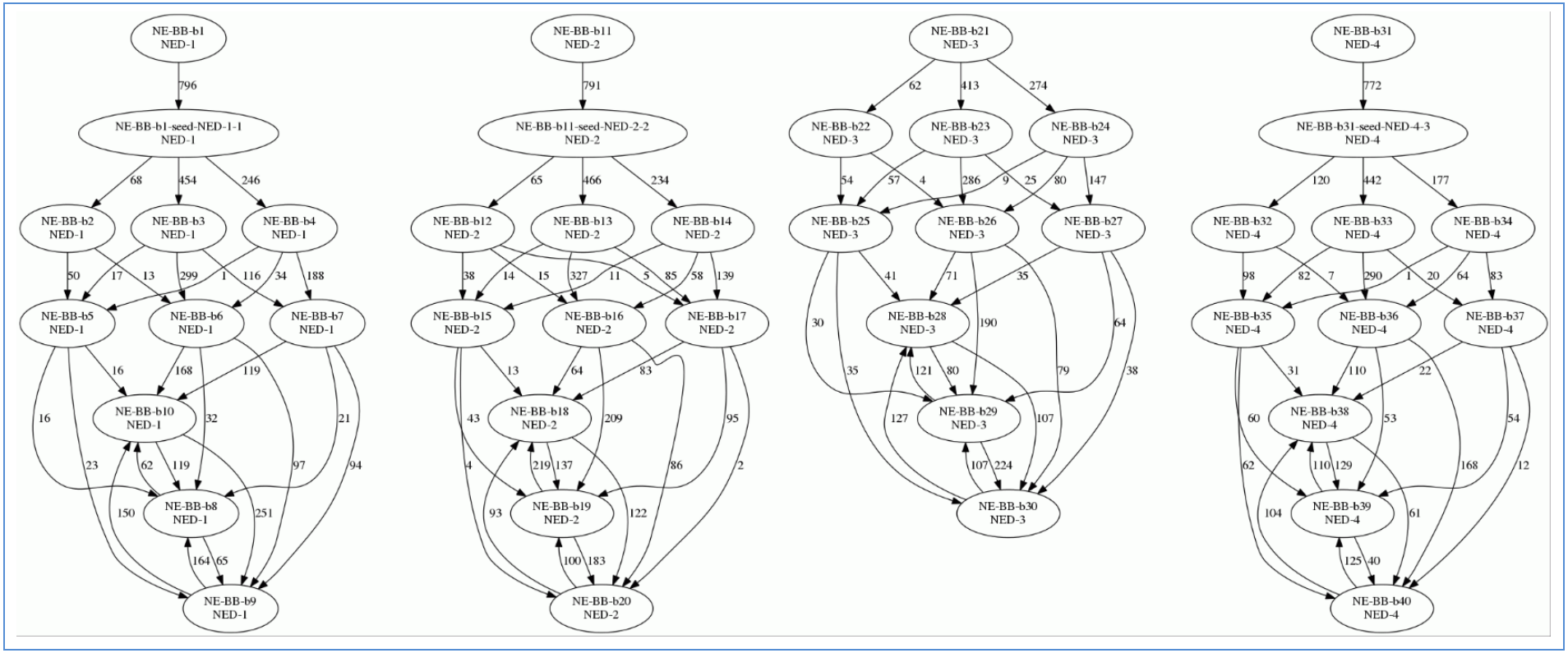


Figure 17-1. Visualizations of the paths followed in the adaptive assessment 2017 of writing Dutch for vmbo-bb

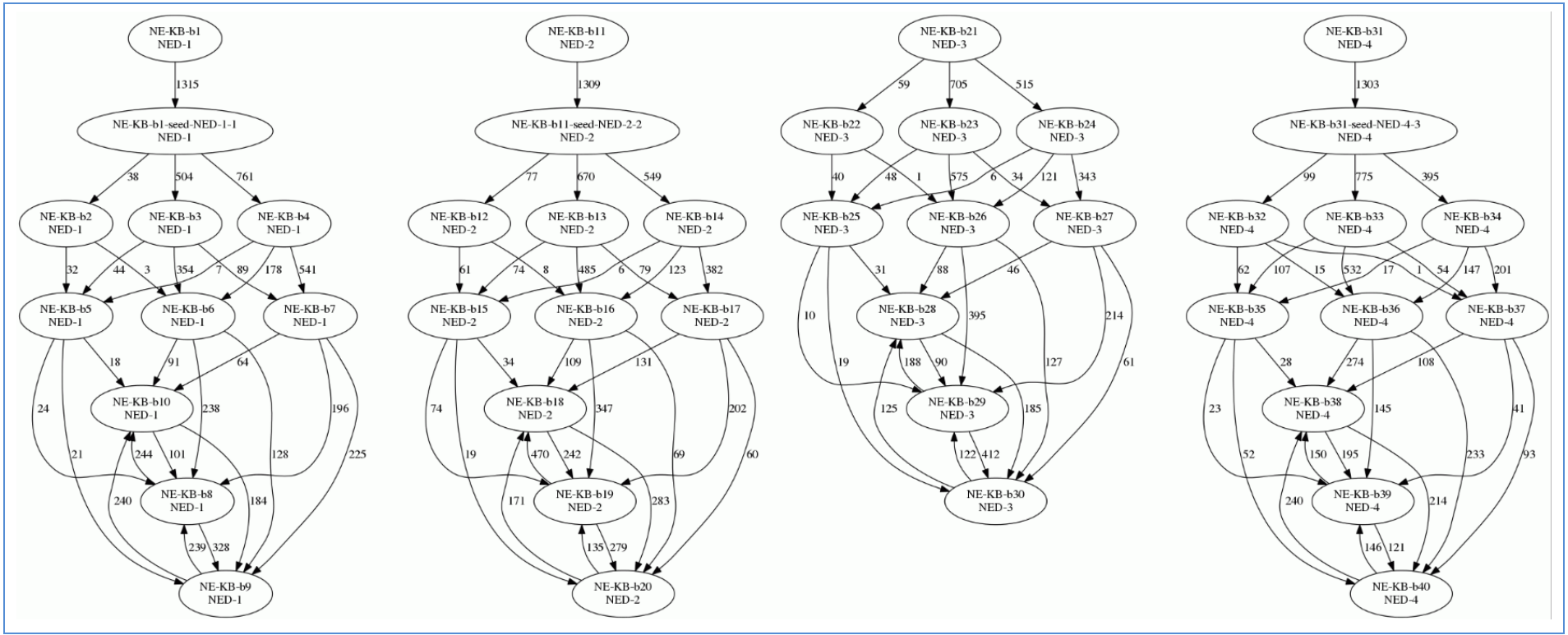


Figure 17-2. Visualizations of the paths followed in the adaptive assessment 2017 of writing Dutch for vmbo-kb

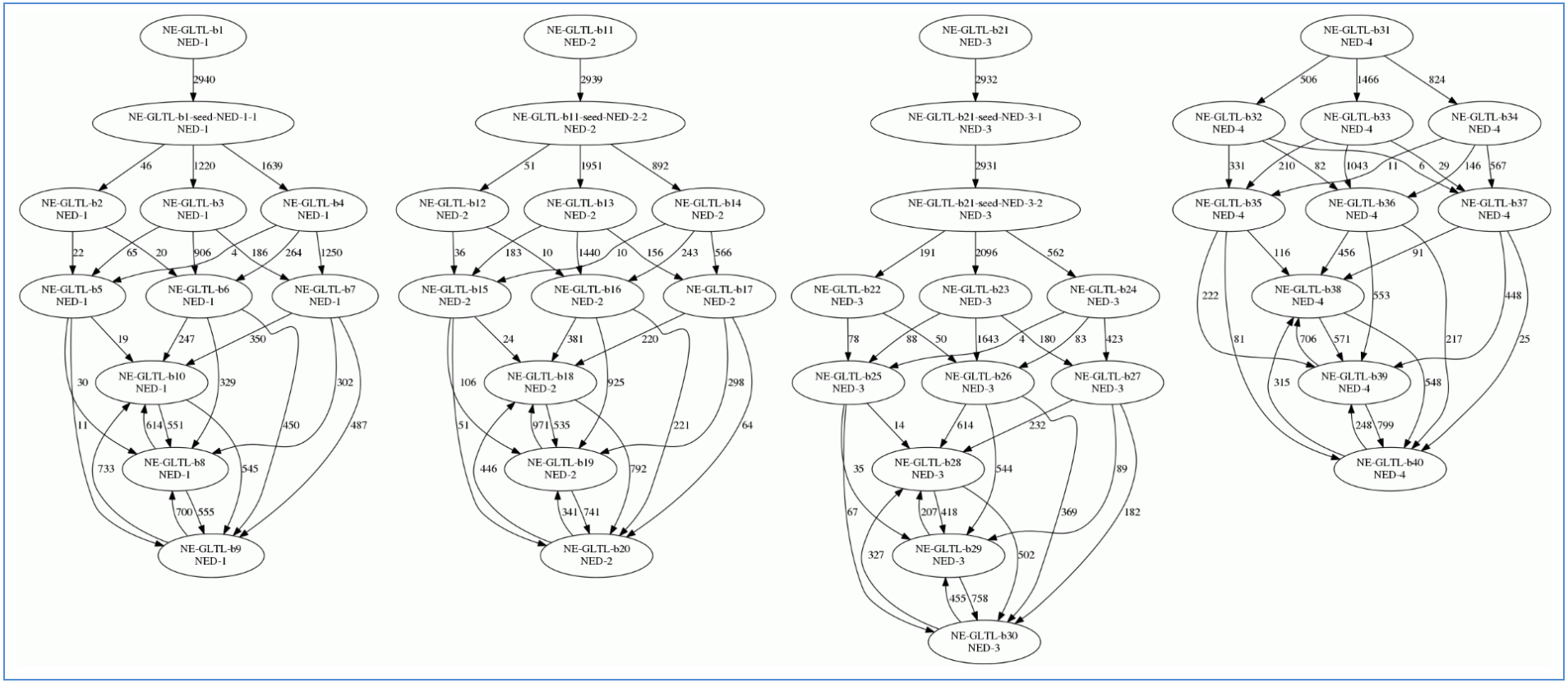


Figure 17-3. Visualizations of the paths followed in the adaptive assessment 2017 of writing Dutch for vmbo-gt

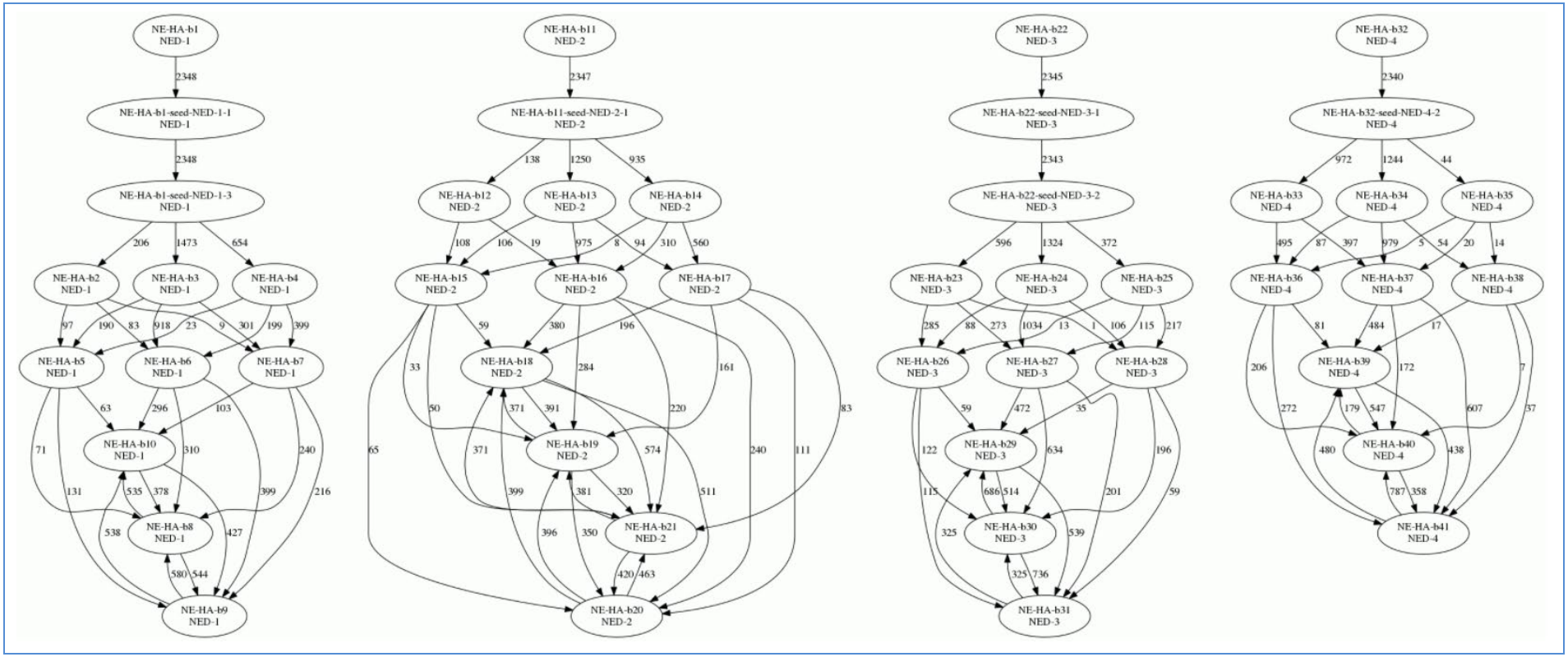


Figure 17-4. Visualizations of the paths followed in the adaptive assessment 2017 of writing Dutch for havo

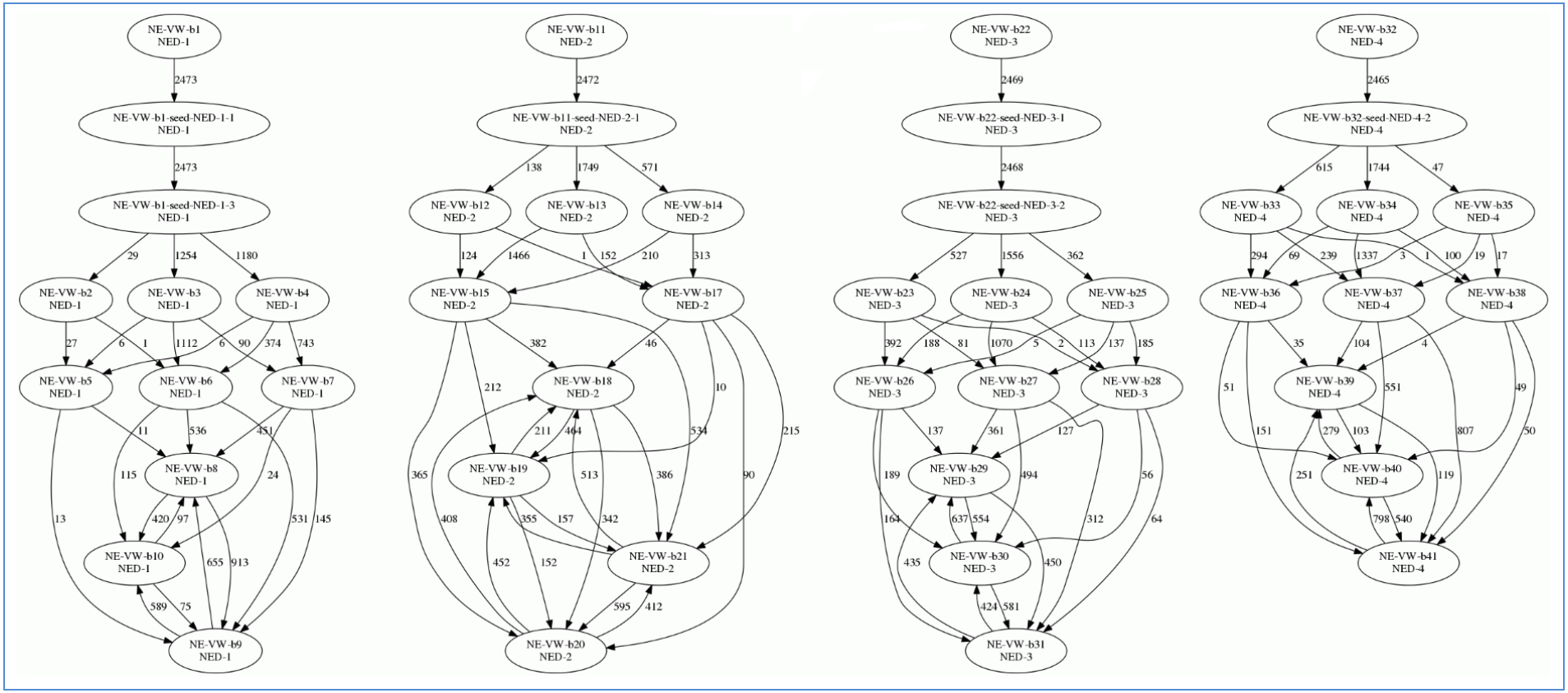


Figure 17-5. Visualizations of the paths followed in the adaptive assessment 2017 of writing Dutch for vwo

17.3.2 Visualizations of the paths followed in the adaptive assessment 2017 of writing English

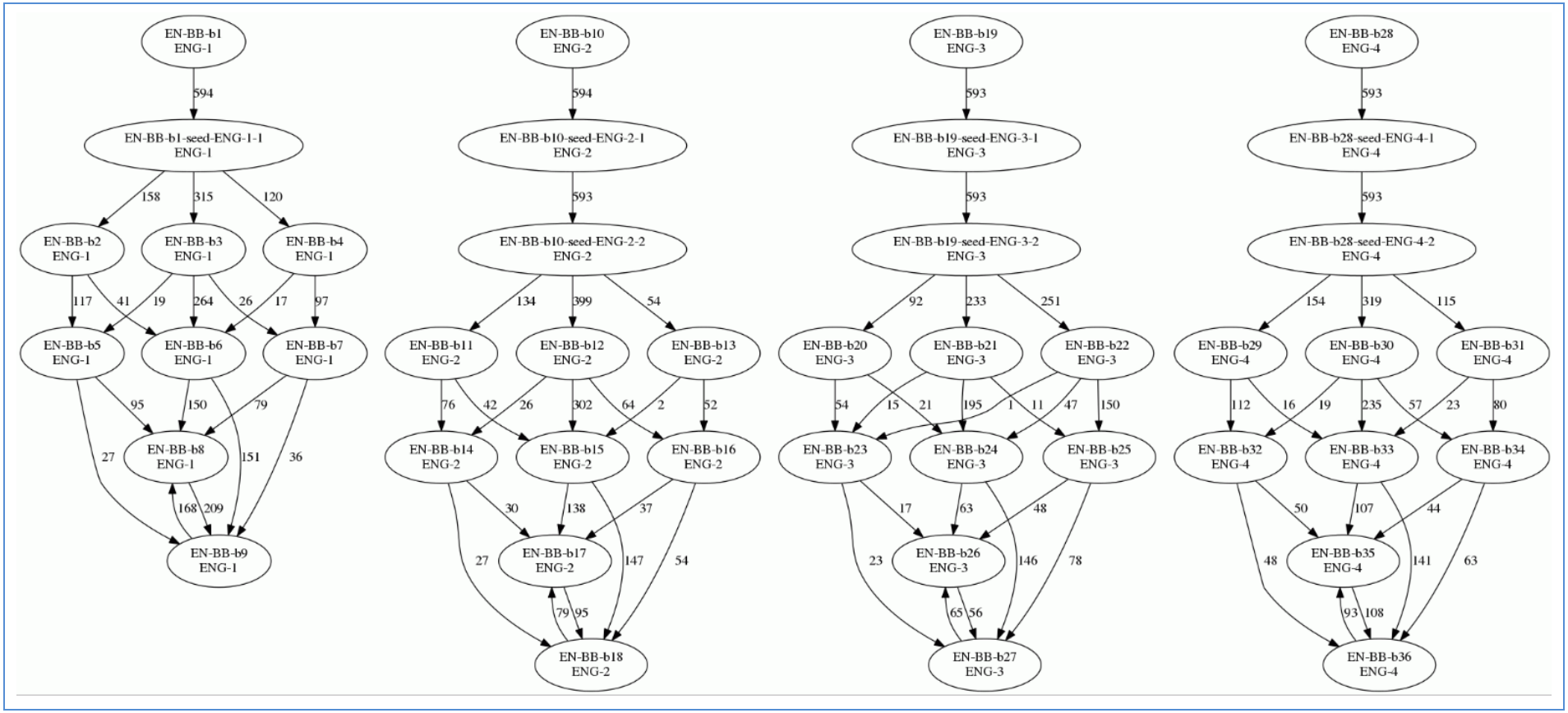


Figure 17-6. Visualizations of the paths followed in the adaptive assessment 2017 of writing English for vmbo-bb

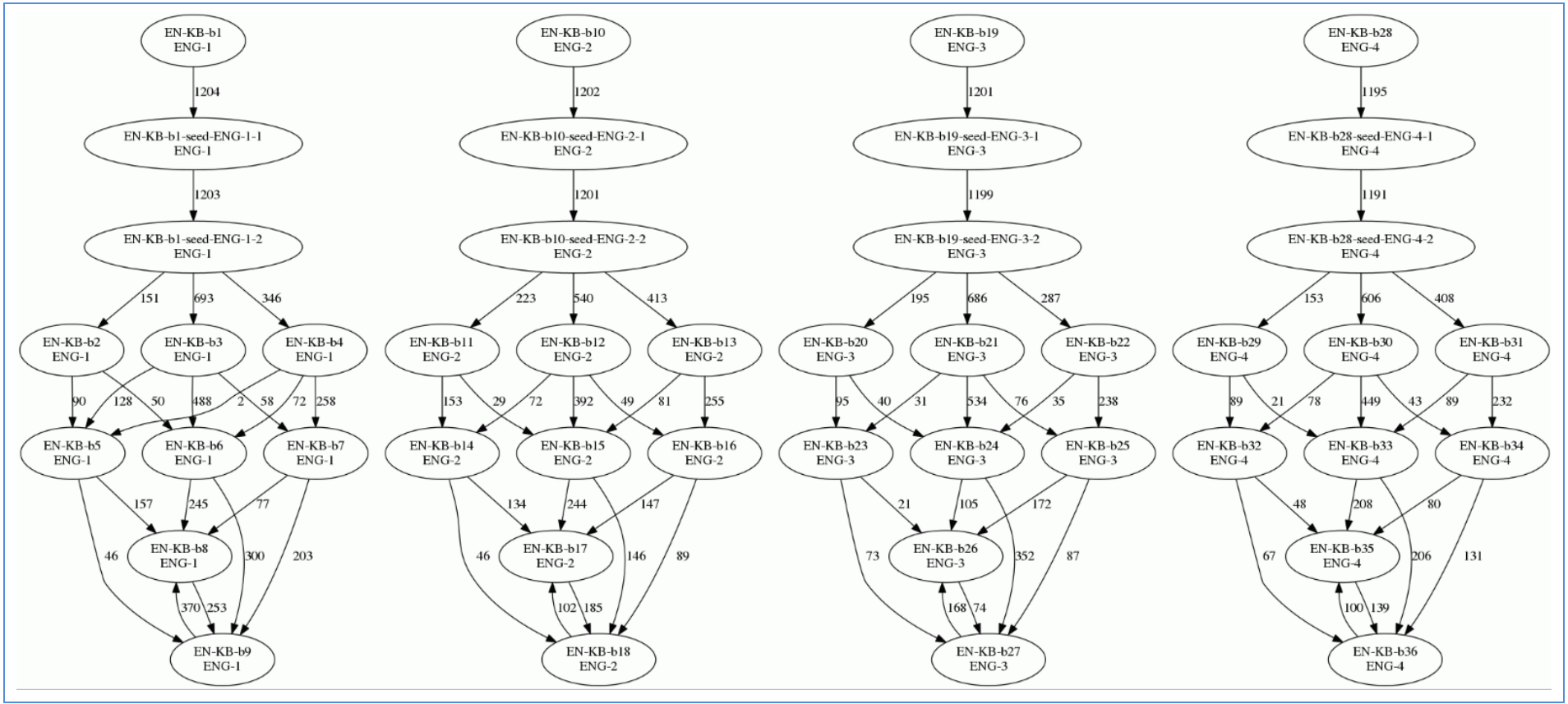


Figure 17-7. Visualizations of the paths followed in the adaptive assessment 2017 of writing English for vmbo-kb

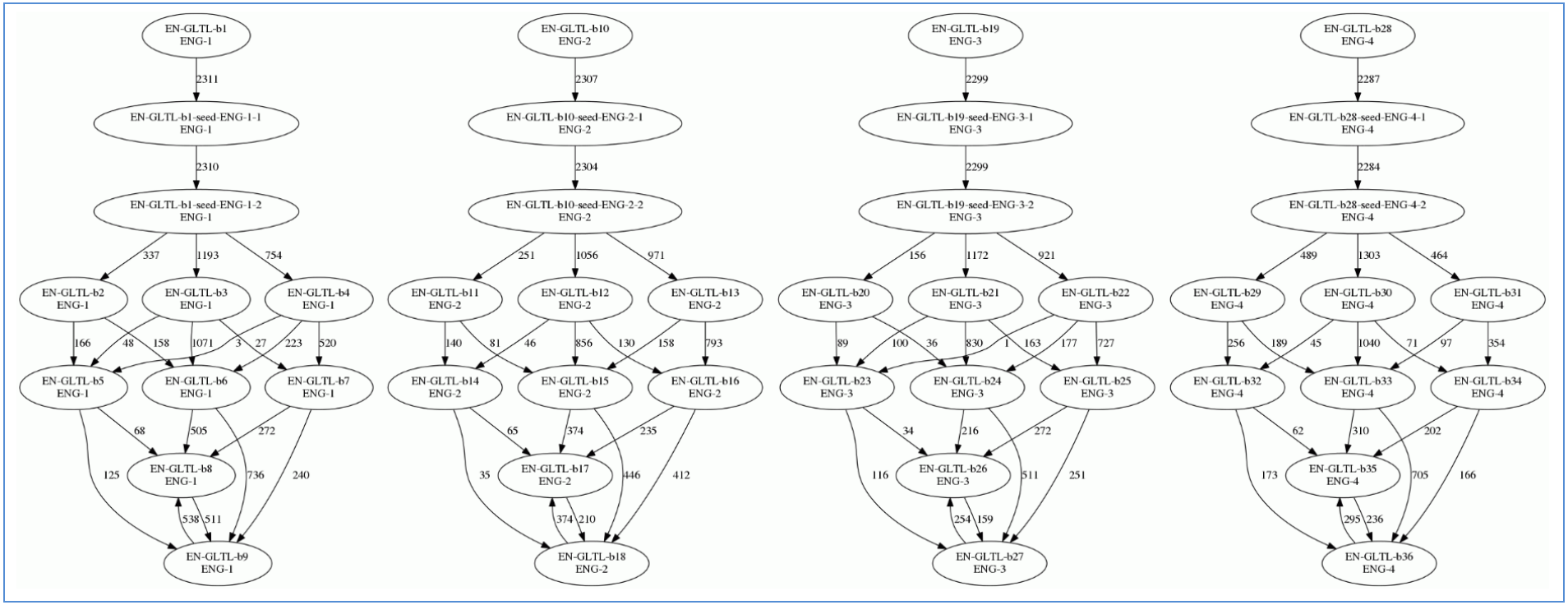


Figure 17-8. Visualizations of the paths followed in the adaptive assessment 2017 of writing English for vmbo-gt

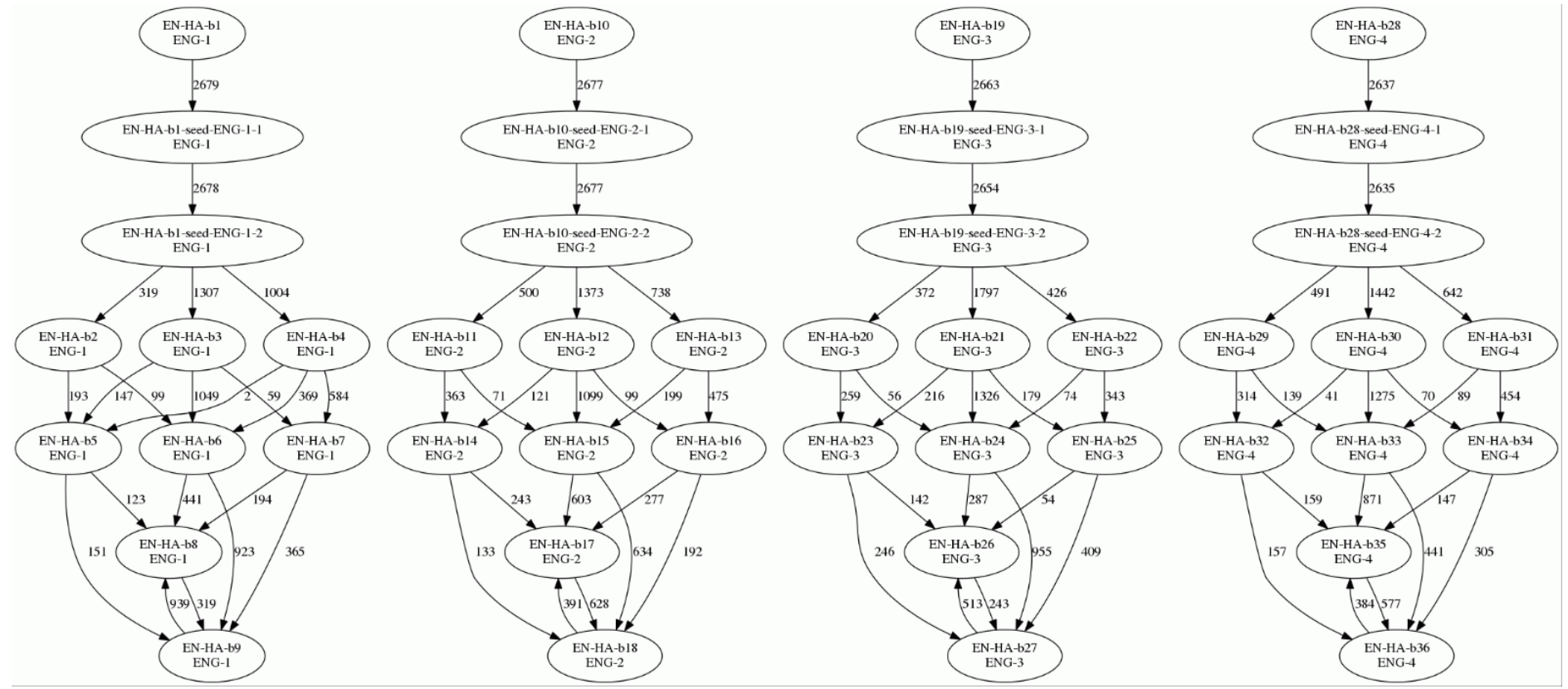


Figure 17-9. Visualizations of the paths followed in the adaptive assessment 2017 of writing English for havo

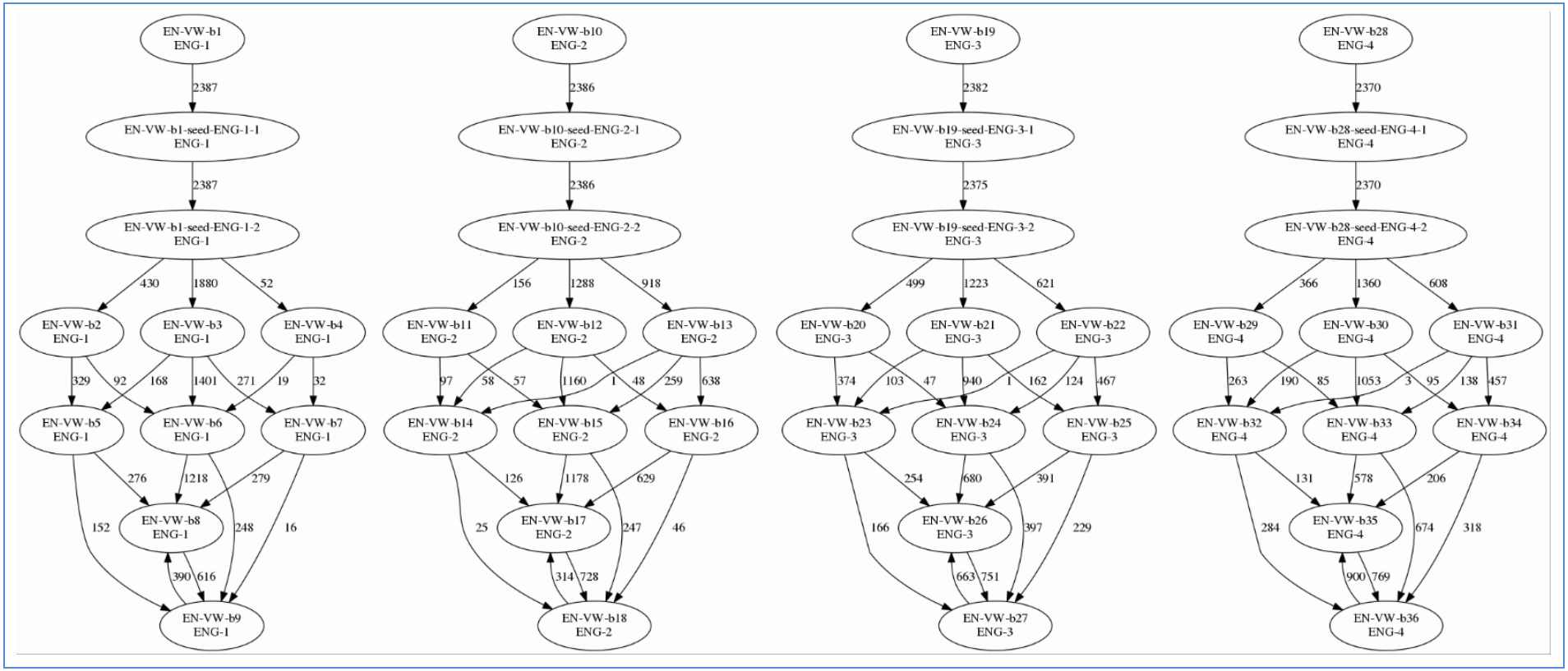


Figure 17-10. Visualizations of the paths followed in the adaptive assessment 2017 of writing English for vwo

17.3.3 Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics

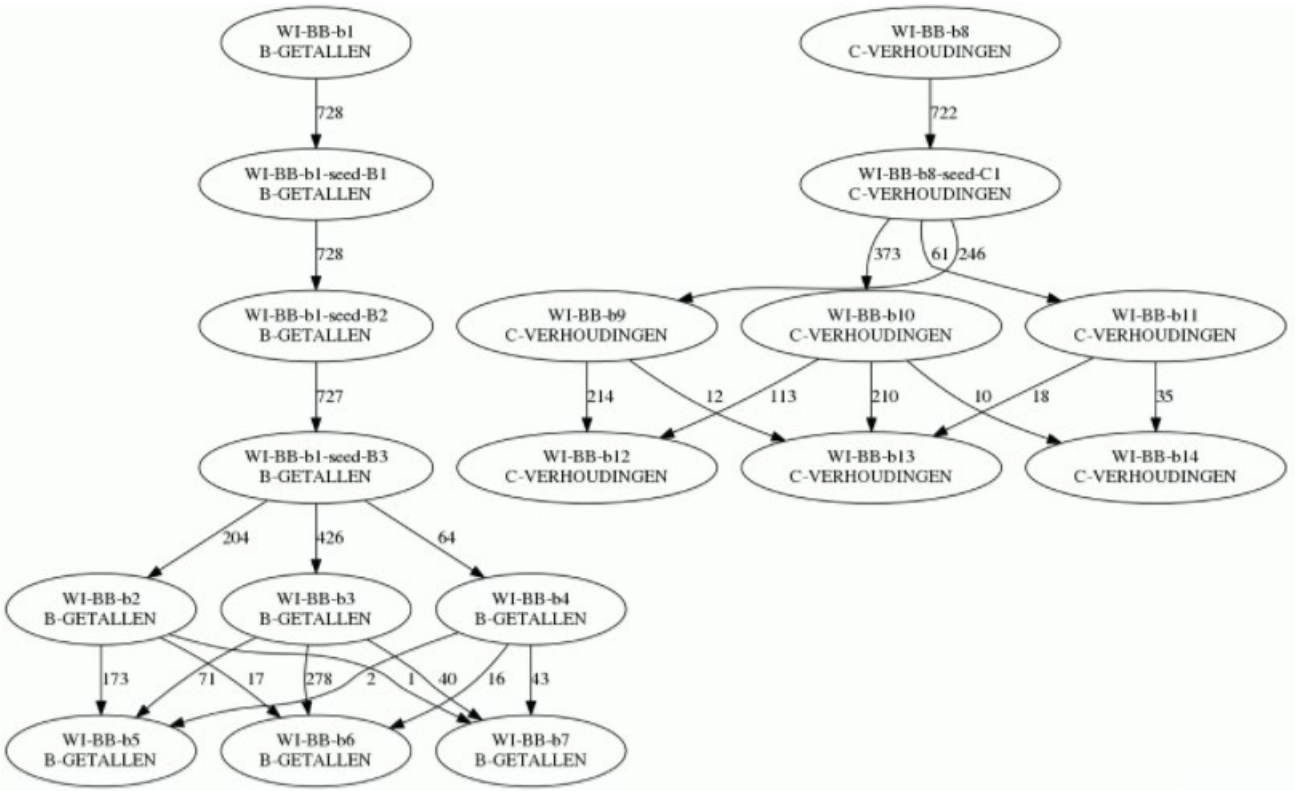


Figure 17-11. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains B and C, for vmbo-bb

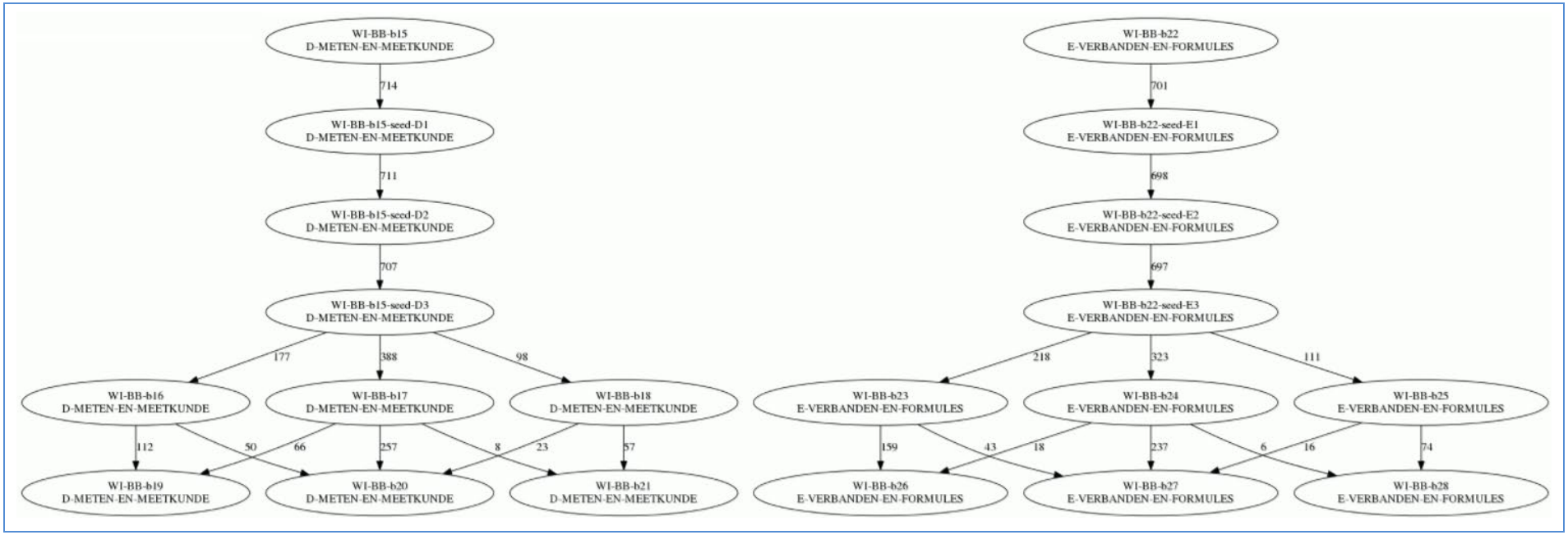


Figure 17-12. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains D and E, for vmbo-bb

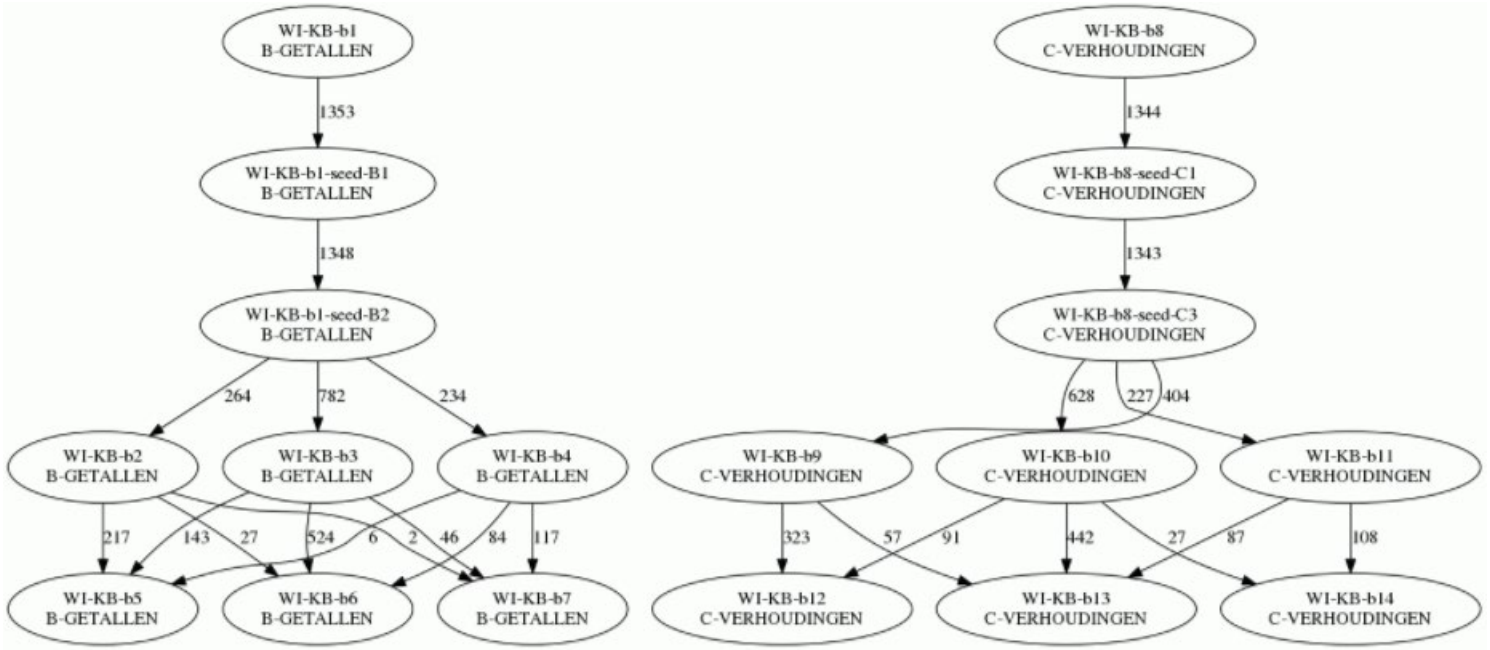


Figure 17-13. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains B and C, for vmbo-kb

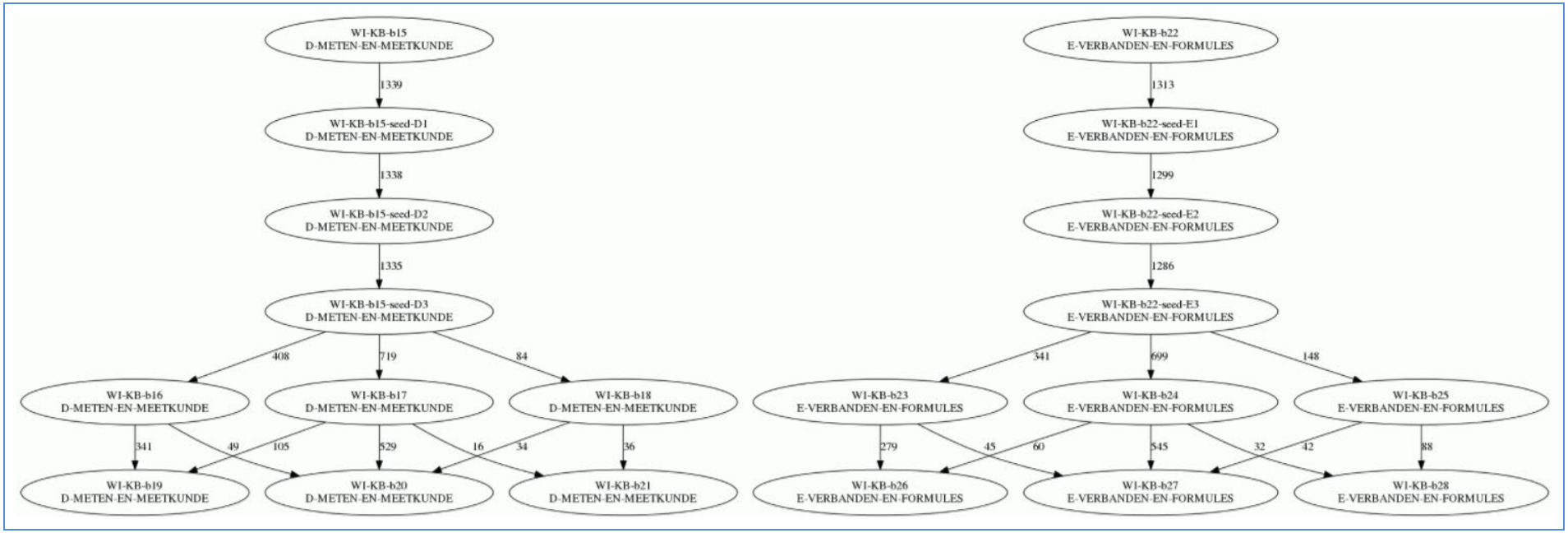


Figure 17-14. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains D and E, for vmbo-kb

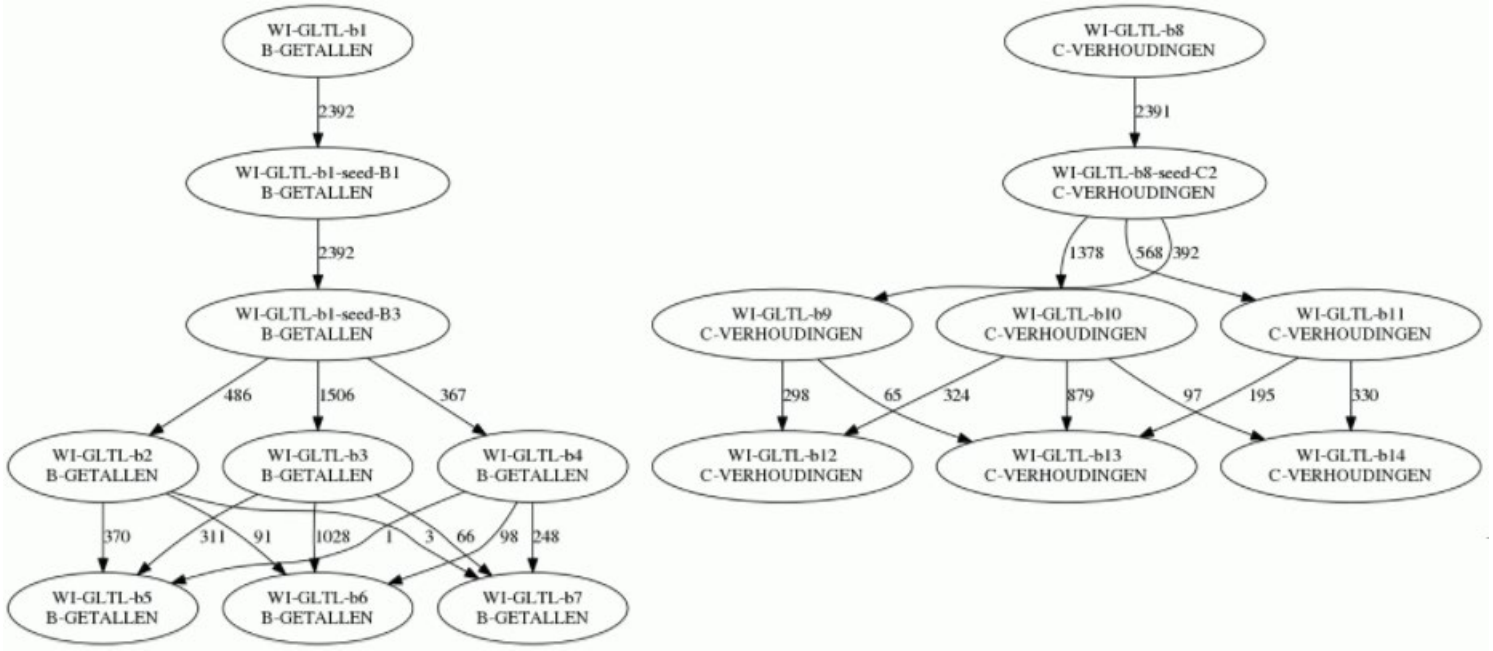


Figure 17-15. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains B and C, for vmbo-gt

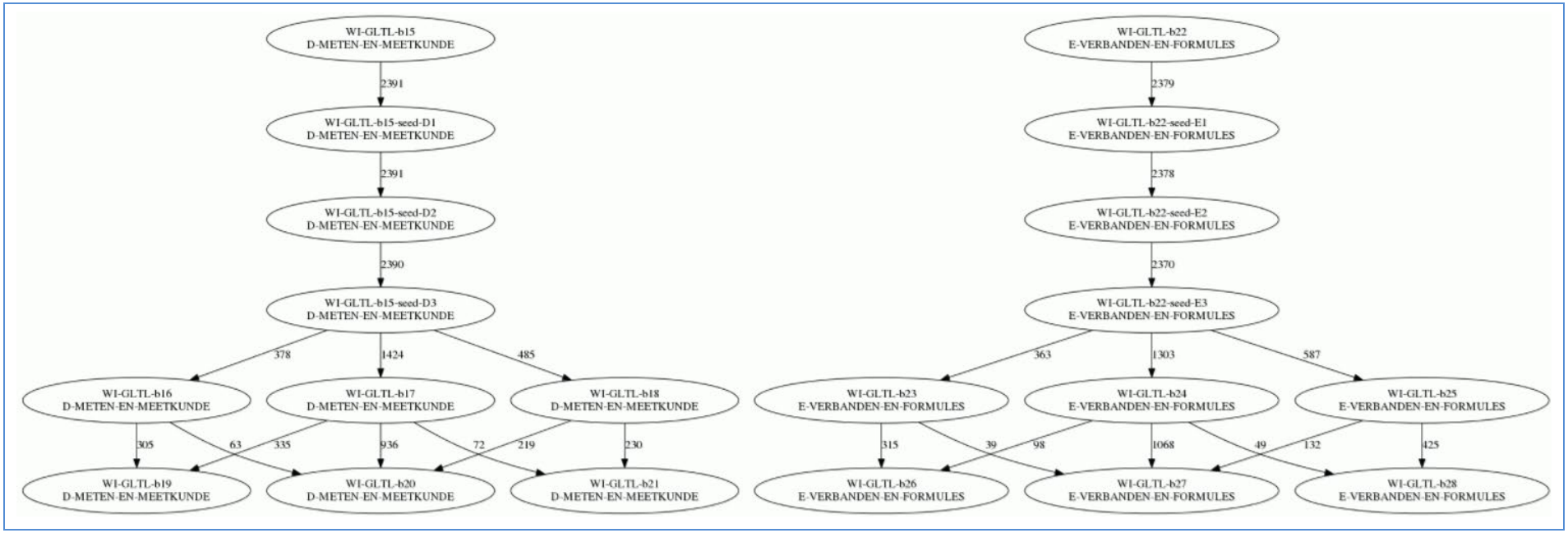


Figure 17-16. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains D and E, for vmbo-gt

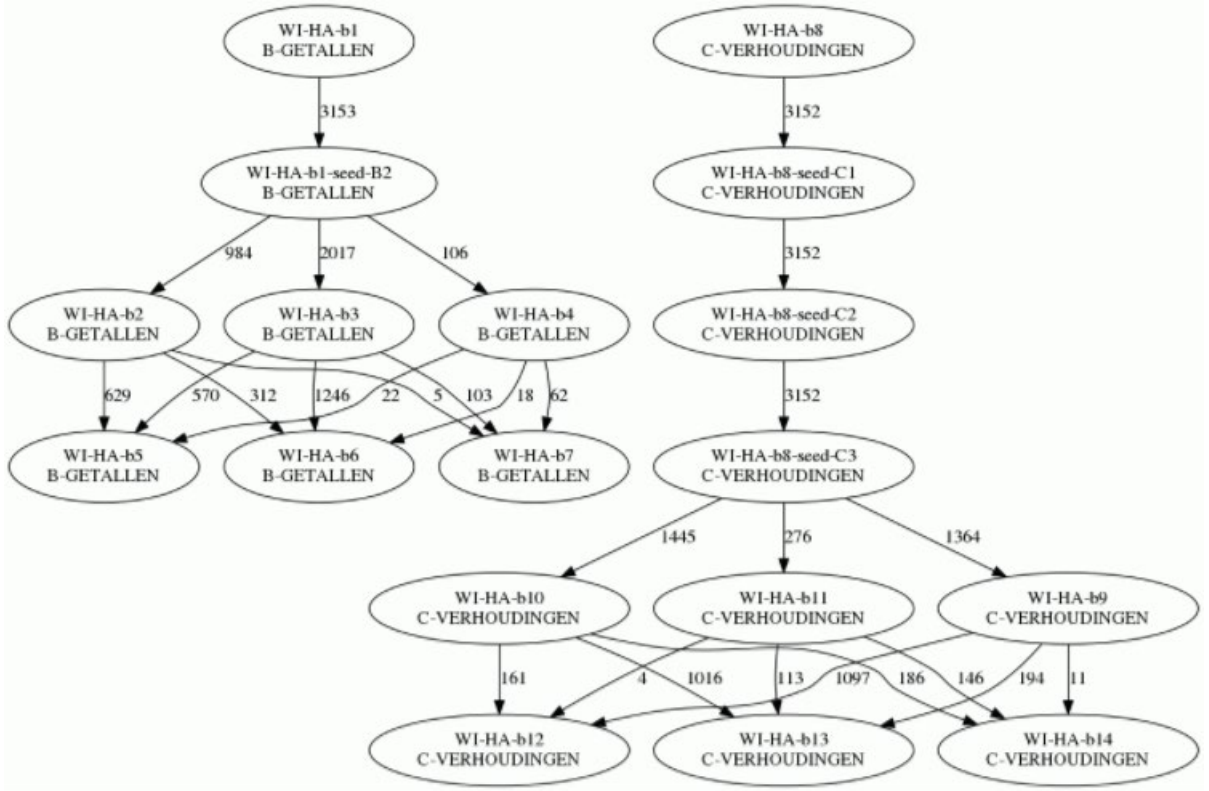


Figure 17-17. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains B and C, for havo

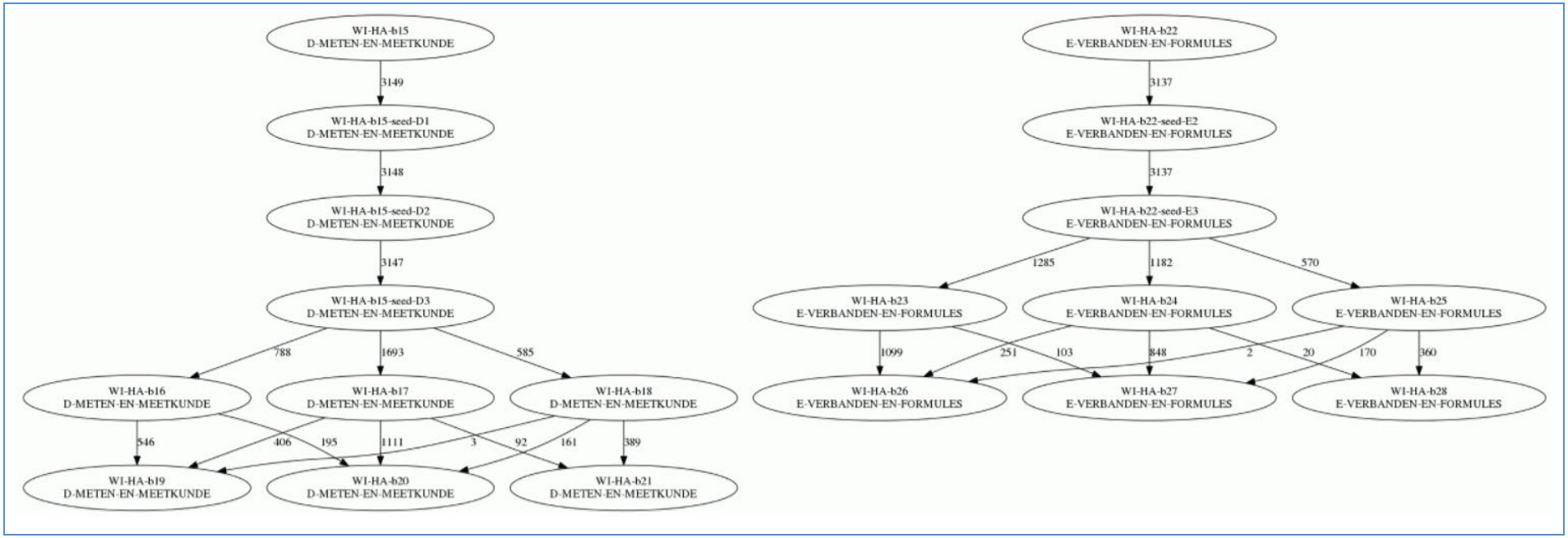


Figure 17-18. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains D and E, for havo

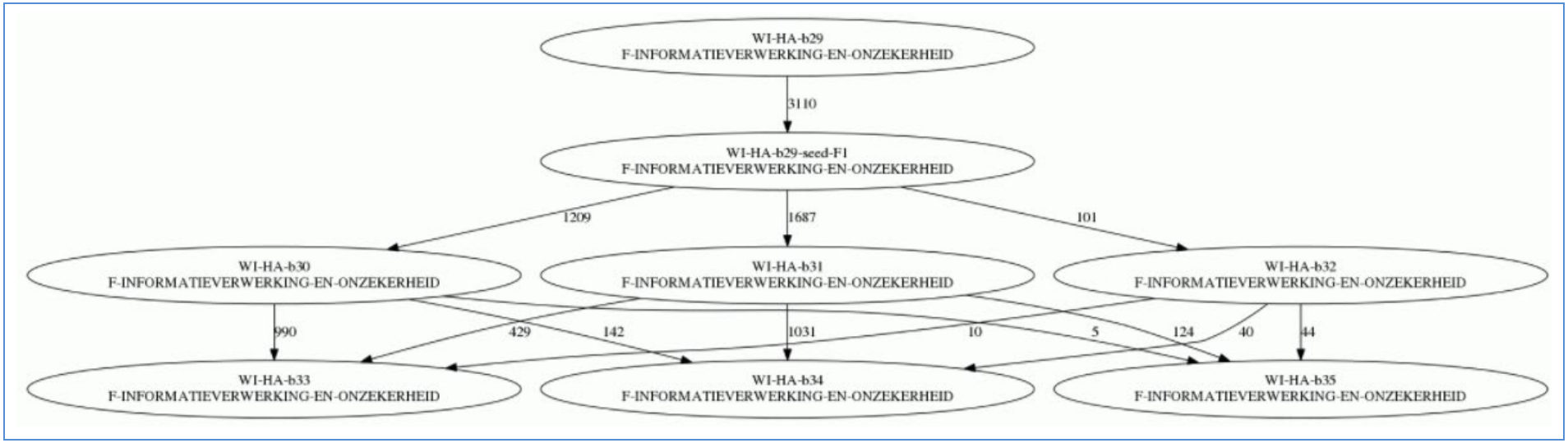


Figure 17-19. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domain F, for havo

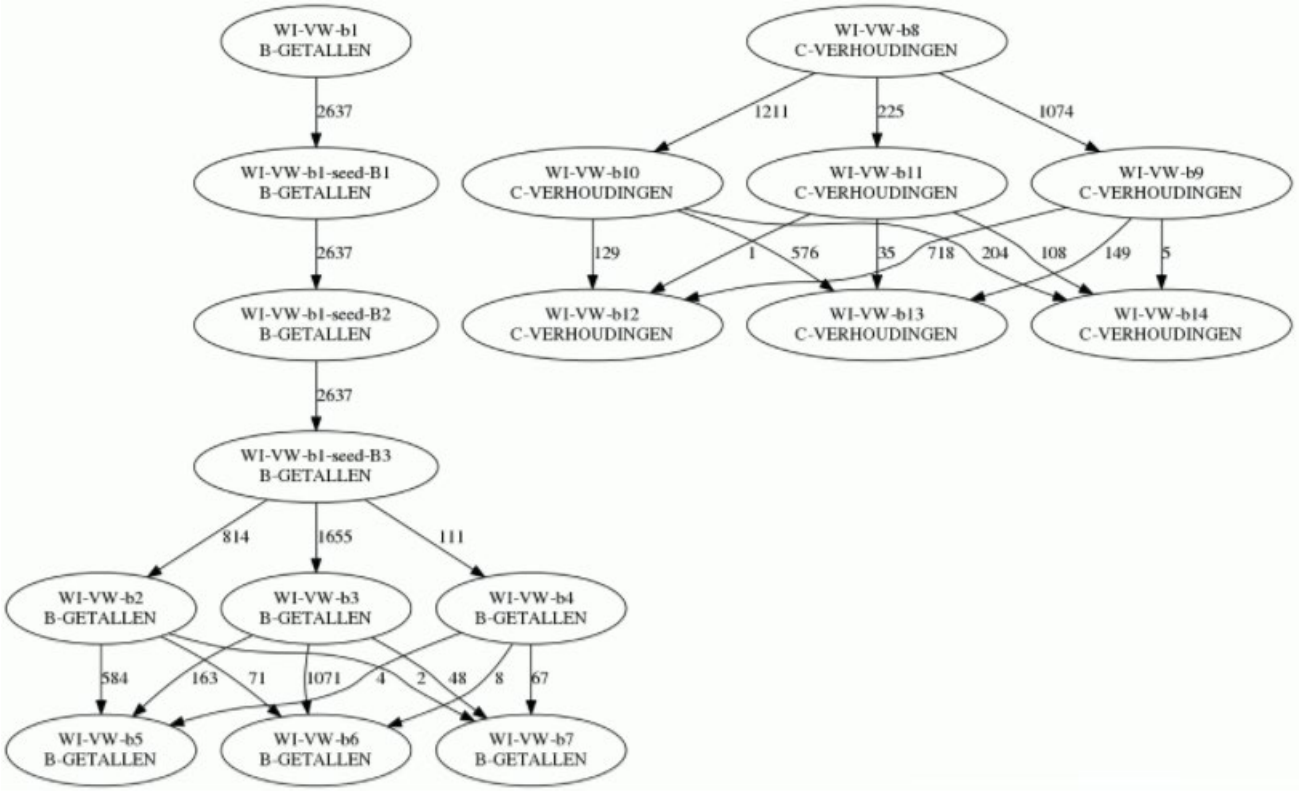


Figure 17-20. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains B and C, for vwo

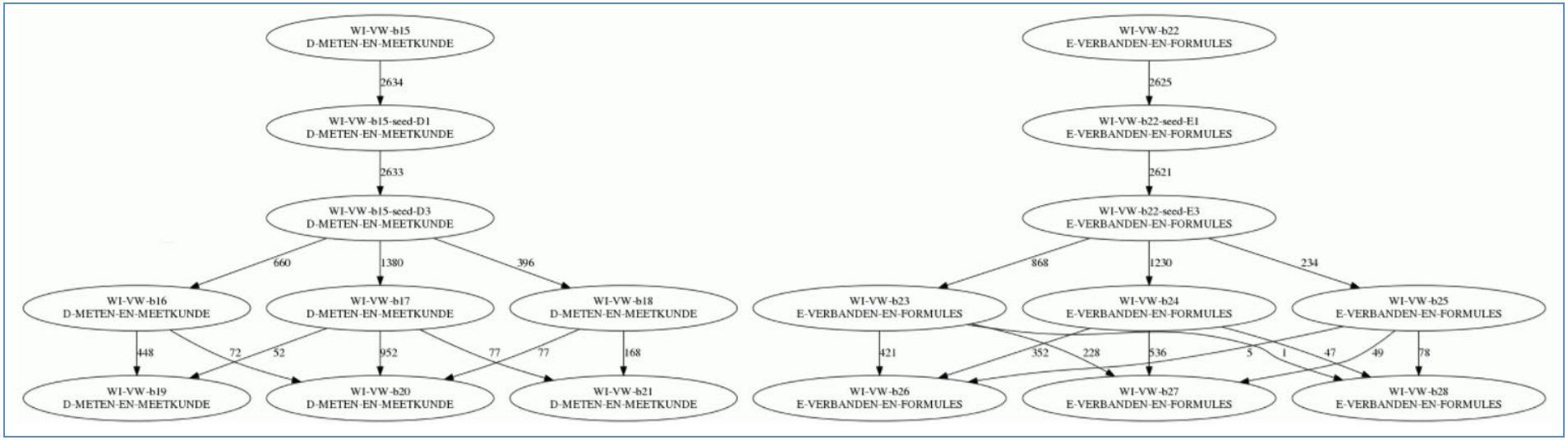


Figure 17-21. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domains D and E, for vwo

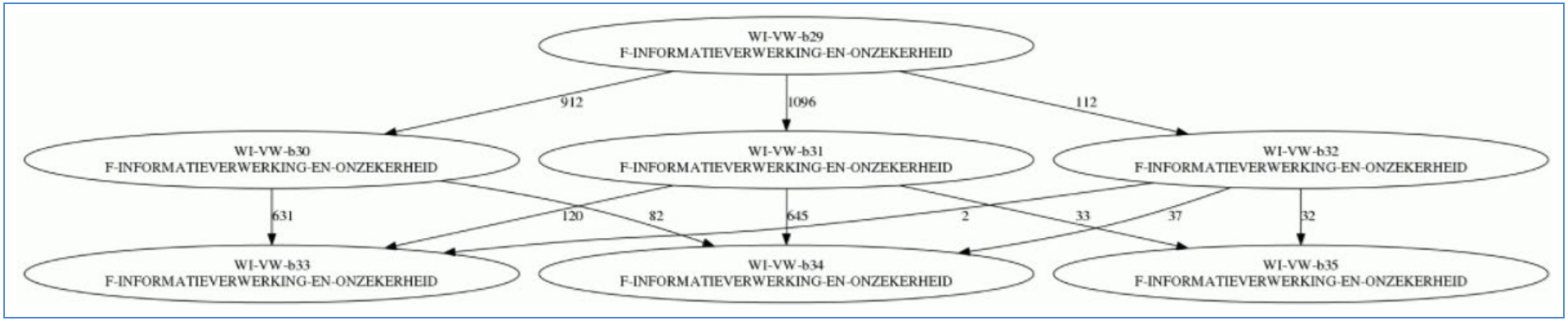


Figure 17-22. Visualizations of the paths followed in the adaptive assessment 2017 of Mathematics, domain F, for vwo

Cito
Amsterdamseweg 13
6814 CM Arnhem
Postbus 1034
6801 MG Arnhem
T (026) 352 11 11

Fotografie: Gijs Versteeg