# E-testing and computer-based assessment

CIDREE Yearbook **2024**

**Authors:**

Lieselot Vandenhoute, Bram Born, Griet Van den Eynde, Soetkin De Knijf, Dries Vrijsen, Kristóf Velkey, Fási Andrea, Emer Delaney, Adrian O'Flaherty, Rachel Perkins, Aidan Clerkin, Rachel Cunningham, Osman Buleshkaj, Fatmir Elezi, Selim Mehmeti, Dr. Philipp Sonnleitner, Steve Bernard, Dr. Sonja Ugen, Max van der Velde, Bernard Veldkamp, Jos Keuning, Max van der Velde, Bernard Veldkamp, Jos Keuning, Kevin Steinman, Katarina Aleksić, Katarina Glišić, Branislav Ranđelović, Ana Radović, Andrejka Slavec Gornik, Karin Hector-Stahre.

**Editors:**

Branislav Ranđelović, Elizabeta Karalić, Katarina Aleksić, Danijela Đukić

**CIDREE national coordinator:**

Danijela Đukić

**Graphic design, layout and cover:**

Institute for Education Quality and Evaluation

**Published online:**

https://ceo.edu.rs/wp-content/uploads/2024/11/CIDREE_Yearbook_2024.pdf

**How to cite this publication:**

Ranđelović B., Karalić E., Aleksić K., Đukić D. (Ed.) (2024). E-testing and computer-based assessment. CIDREE Yearbook 2024. Belgrade: Institute for Education Quality and Evaluation.

© CIDREE: Consortium of Institutions for Development and Research in Education in Europe CIDREE is a network of educational organizations involved in curriculum development and/or educational research, set up in 1991 to establish closer working relationships at a European level. All rights reserved. No part of the publication may be reproduced, stored in an information storage and retrieval system, or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, without prior permission in writing from the publisher.

ISBN – 978-86-6072-153-4

# Contents

# Editorial
## Introduction

The theme of the CIDREE Yearbook 2024 is E-testing and computer-based assessment. We are living in the age of digitalization and wide use of computers and digital devices. Testing in schools and any form of computer assessment seems to be one of the most important processes in contemporary education (Al-Maawali, W., & Al Rushaidi, I., 2024), (Ortiz-Lopez, A., Olmos-Miguelanez, S., & Sanchez-Prieto, J. C. , 2024). Additionally, we have increased need and we put a large focus in measuring student outcomes. This Yearbook provides an insight into E-testing and computer-based assessment from several European educational systems. Throughout ten articles, authors from diverse professional backgrounds and specializations offer different perspectives on promoting E-testing and computer-based assessment in wide education contexts. They share experiences and aims of their own educational systems. Throughout those articles, there is a clear focus on the need to engage digital technology and e-testing tools in assessment process (Randjelovic, B., Aleksic, K., Stanojevic, D., 2020), (Shute, V. J. & Rahimi. S., 2017).

The articles explore the diverse challenges faced by educational stakeholders in different educational systems, different school environments and various digital capacities, aiming to share knowledge and experiences, in order to increase the quality of the educational processes and to prepare the students for the challenging world of the twenty-first century.

## A glimpse into the Yearbook

Authors of ten European educational systems have contributed articles to the Yearbook 2024: Flanders-Belgium, Hungary, Ireland, Kosovo, Luxembourg, Netherlands, Norway, Serbia, Slovenia and Sweden. Below you will find a short introduction to each of the articles that are included in the CIDREE Yearbook 2024:

## FLANDERS - BELGIUM

HIGH-QUALITY SUMMATIVE E-TESTING AT THE EXAMINATION BOARD FOR SECONDARY EDUCATION IN FLANDERS

The examination program that participants must complete consists mainly of digital exams to be taken at the professionally equipped examination center in Brussels. More than 45,000 exams are conducted on an annual basis. In 2012, the Examination Board started the transition from written exams to digital exams and has since been permanently committed to improvements through technological developments. E-testing has made the process of conducting exams more efficient. They preferably consist of close-ended questions, which enables automatic correction. Developing these type of ques¬tions requires specific expertise and adequate staff training, which are crucial for e-testing. All digital exams are automatically graded. This paper ends with a critical view of e-testing on the Flemish Examination Board and a look at future developments.

## HUNGARY

FROM PAPER TO ONLINE ASSESSMENT

Digital assessments have been in place since 2022. Authors share experiences on the reasons behind the need for and the process of digital transformation, many challenges during the planning and implementation. In Hungary, there are six assessment areas (Mathematics, Reading Comprehension, Science, Foreign Languages (English, German), History, IT) from this school year onwards. Digital assessments will be carried out in eight grade, which also brings significant challenges in terms of the organization and implementation of the assessments.

## IRELAND

COMPUTER-BASED TESTING IN IRELAND, 2005-2024: CHALLENGES, LESSONS LEARNED, AND FUTURE POSSIBILITIES

This chapter traces the use of computer-based testing in schools in two contexts: international large-scale assessments and national standardized testing. Learning gleaned from participation in computer-based ILSAs informed the development of a bespoke online platform for administering and scoring standardized tests ([ERC DOTS]). Conversely, knowledge about schools' use of ERC DOTS has informed Irish approaches in subsequent ILSA cycles. Across ILSAs and standardized testing, recurrent challenges of computer-based testing are identified, including wide variation in: (i) education technology infrastructure in schools; (ii) students' prior experiences using this for schoolwork. Although common to primary and post-primary settings, infrastructural challenges were most apparent at primary level while differences in students' experience were most pronounced below Grade 3.

## KOSOVO

E-TESTING AND COMPUTER-BASED ASSESSMENT IN KOSOVO

In recent years, e-testing and computer-based assessments have been used during international assessments such as PISA and TIMSS. These implementations faced challenges related to infra¬structure, access, and preparation. Additionally, various schools have adopted on¬line platforms for computer-based assessments, more about the circumstances created in the conditions of the COVID 19 pandemic. However, these efforts currently lack cohesive national policy guidance. This study aims to investigate the current state of e-testing and computer-based assessments, focusing on infrastructure, access to technology, training provisions, and an analysis of the benefits, challenges, and potential for application.

## LUXEMBOURG

EMERGING TRENDS IN E-ASSESSMENT: INSIGHTS FROM OASYS AND THE IMPACT OF GENERATIVE ITEM MODELS

This chapter provides a comprehensive exploration of e-assessment, delving into its multifaceted advantages and challenges. It begins by tracing the developmental path and critical insights gathered from the utilization of the OASYS (Online-Assessment SYStem) assessment platform, a cornerstone of educational practices. Subsequently, the focus shifts towards a critical aspect prevalent in all e-assessments: the creation of test content. Regardless of the intended application, whether it is used for adaptive testing, formative assessments, or summative evaluations, the necessity of robust and psychometrically sound test content remains paramount. Within this context, the chapter illustrates the innovative approaches adopted by the Luxembourg Centre for Educational Testing (LUCET) in addressing this challenge. Specifically, it highlights the implementation of template-based, generative item models in a large-scale mathematics assessment conducted nationwide. Furthermore, the chapter explores the growing interest in generative artificial intelligence (AI) and its potential implications in this context. Through a nuanced examination of these themes, this chapter offers valuable insights into the current trends and future directions of e- assessment.

## NETHERLANDS

### THE FRAMEWORK AND DEVELOPMENT OF SERDA: SPEECH ENABLED READING FLUENCY ASSESSMENT FOR DUTCH

The importance of reading for educational, vocational and societal life cannot be understated. Nonetheless, recent large-scale studies reveal that the reading comprehension of students has declined globally, and specifically in the Netherlands. Developing fluent reading skills allows children to read quickly, accurately and with proper expression, which is fundamental to becoming a good reader. To monitor this development, teachers need to assess fluency on a regular basis. However, fluency assessment is currently time-consuming for teachers, provides limited information, and neglects prosody assessment. This chapter presents a framework for, and the development of, a digital automatic fluency assessment tool for early primary education that overcomes current issues through incorporating Automatic Speech Recognition (ASR): the Speech Enabled Reading Diagnostics App (SERDA). The results provide usability, validity and reliability evidence for SERDA's speed and accuracy measures. Furthermore, SERDA reduces the testing burden placed on teachers, increases the information gained, and facilitates prosody assessment.

## NORWAY

### NORWAY'S 2024 GENERATIVE AI JOURNEY IN SCHOOLS AND ASSESSMENTS: RESPONDING TO A CALL FOR GUIDANCE ON NEWER DIGITAL TECHNOLOGIES IN NORWAY'S SECONDARY SCHOOLS

This article presents an exploration of both the challenges and opportunities, teachers and pupils are experiencing since Large Language Models' (LLMs') sudden entry into the classroom two years ago. By establishing new guidance, the Directorate's role as an advising authority is highlighted in the article's first section, along with the challenges of advising the education sector in the early stages of ongoing technological revolutions. The second section details the Directorate's response to generative AI's prospective impact on centralized, written assessments. The article concludes by unpacking the recent digitalization of the high-stakes, secondary English written assessment.

## SERBIA

### E-TESTING AND COMPUTER-BASED ASSESSMENT IN SERBIA

This paper offers an overview of the achievements, current state, and trends in Serbian education concerning e-testing. It highlights examples from various educational segments and levels, including primary education (such as International Large-Scale Assessments like TIMSS, PIRLS, and the Final Exam Field Trial), secondary education (including PISA), and e-testing in adult education. The educational system is highly focused on digitalization, aiming to ensure that all students develop digital competence. Current outcomes reflect progress toward this objective, and the entire educational system is eagerly awaiting the full implementation of e-testing procedures.

## SLOVENIA

SLOVENIA’S TRANSITION TO E-MARKING AT NATIONAL EXAMINATIONS

National Examinations Centre administers external assessments, including the National Assessment at the end of grades 6 and 9 in primary education and Matura examinations at the end of secondary education. In recent years, we have completed the transition from paper-based marking to electronic marking of examinations at both levels. The article describes the experience with the implementation of e-marking, the activities that preceded the introduction of e-marking, the challenges that are faced during the preparation and implementation phases, and the benefits that resulted from the implementation of e-marking. A pilot e-testing project from 2021 is also briefly addressed, as well as the challenges that lie ahead.

## SWEDEN

THE EXPERIENCE OF DEVELOPING AND LAUNCHING E-TESTING ON A NATIONAL LEVEL

Digital national tests are introduced on a national level in the spring of 2024. In this paper, the process of developing and gradually introducing e-testing is described. The article describes several challenges in starting up a system for digital national tests available for all students within the last year of compulsory school, grade 9, upper secondary school, and adult education at the upper secondary level. Particular attention is given to the adaptation of the user interface in the digital assessment platform. The interface has been comprehensively adapted to comply with the legislation for accessibility. The task of increasing the preparedness of schools and school organizers in implementing the use of digital national tests is described as well as the challenges to implementing such tests in a decentralized school system. Another issue addressed in the article is the gradual development of external scoring and assessment, which will entail the recruitment of over 3,000 certified teachers.

## Conclusion

This introduction highlights that one of the most important tools in education of the 21st Century is digitalization of assessments and e-testing. To feel more motivated to do something at assessments and to have a better score, students generally need to use modern tools and to feel excited about it. Many examples from various educational systems are shown, proving that this process is ongoing in almost all systems in Europe. We hope that the work described in this CIDREE Yearbook 2024, and the insights shared by our authors into their fields of expertise will hopefully provide grounds for fruitful discussion and reflection. In particular, it is hoped that all policy makers, researchers and practitioners across Europe and beyond will find this Yearbook a useful resource to inform their communities and educational systems, regarding e-assessments.

## References

Al-Maawali, W., & Al Rushaidi, I. (2024). Teachers' Perspectives on Affordances and Challenges of Technology for Reliable and Valid Online Testing: Learning the Lessons from the COVID-19 Pandemic. Ubiquitous Learning: An International Journal, 17(1).

Ortiz-Lopez, A., Olmos-Miguelanez, S., & Sanchez-Prieto, J. C. (2024). Toward a new educational reality: A mapping review of the role of e-assessment in the new digital context. Education and Information Technologies, 29(6), 7053-7080.

Randjelovic, B., Aleksic, K., Stanojevic, D. (2020). Online self-assessment as preparation for final exam in primary schools – experience from COVID19 crisis, Proceedings of International conference on Applied Internet and Information Technologies AIIT2020, Technical faculty Zrenjanin, 297-300.

Shute, V. J. & Rahimi. S. (2017). Review of computer-based assessment for learning in elementary and secondary education. Journal of computer assisted learning. Vol 33(1), 1–19. https://doi.org/10.1111/jcal.12172

# Foreword

The topic of the 2024 CIDREE yearbook is 'E-testing and computer-based assessment'. The topic is relevant and exciting for many reasons.

In recent years, we have witnessed a huge expansion of technological applications in all areas of life, including in education. The increasing presence and accelerated development of e-learning and e-testing are remarkable. These developments have significant impact on our ways of teaching and learning and our methods of evaluation and quality assurance. It acquires adjustments for students, teachers, schools, curricula and educational systems. Can assessments of students' competencies be successfully implemented in an online environment? Can an online assessment provide high-quality and relevant results of learning outcomes? What are the conditions for a balanced system of online assessment in relation to summative and formative evaluation and student guidance?

Some educational systems are pioneers in the field of e-testing and computer-based assessment, some are still searching for the best and safest models. In the 2024 CIDREE yearbook, we want to explore and discuss how educational systems make decisions when faced with the challenges of e-testing and computer-based assessment. We want to discuss methodological dilemmas, issues of quality assurance, preconditions for implementation and differences and similarities in our experience with e-testing and computer-based assessment.

This yearbook will be launched and discussed at the 2024 CIDREE conference in Serbia. The presence of ten articles of ten different countries shows the relevance of the topic of this year's yearbook. Flanders (Belgium), The Netherlands, Ireland, Kosovo, Luxemburg, Hungary, Norway, Slovenia, Serbia and Sweden have contributed to this yearbook. There are different perspectives and focus areas in the chapters:

- Policy frameworks and preconditions for implementation
- Methodological and quality issues concerning e-testing and computer based assessment
- Challenges and opportunities for teachers and students

The chapters show us that there is a huge amount of issues to consider, but also that there are many shared challenges regarding e-testing and computer based assessment. The yearbook sheds lights on current trends and future directions, and I hope it contributes to good discussions and reflection on the best way forward.
On behalf of all CIDREE members, I would like to thank our Serbian colleagues for taking the initiative for this very interesting and relevant topic for the 2024 Yearbook, as well as for their coordination and editorial work. And, of course, our thanks go out to all the contributing authors. The yearbook reflects the richness of expertise and experience within the CIDREE network.

Enjoy!

**Ingrid Vanhoren**

CIDREE President 2024-2025

Head of Division Qualifications and Curriculum
AHOVOKS Belgium – Flanders

# Flanders-Belgium

# FLANDERS-BELGIUM: BIOGRAPHIES

## Lieselot Vandenhoute

Lieselot Vandenhoute is test development process manager and subject coordinator for geography at the Secondary Education Examination Board in Brussels. There, she is responsible for the geography exams and for ensuring the quality of the exams and exam questions. She began her duties at the Examination Board in 2019, after a career of 16 years as a geography teacher trainer at KU Leuven and Vives University College. Lieselot graduated in 2002 with a master's degree and teaching qualification in geography from Ghent University in Belgium.

## Bram Born

Bram Born is a test expert at Flemish Secondary Education Examination Board in Brussels where he works at the intersection of test development and item analysis to ensure the quality of the examination process. He joined the Examination Board in 2015, taking on the role of overseeing test development and ensuring quality control for the subjects of physics and natural sciences. He began his teaching career in 2003 as a chemistry teacher in Brussels, and later worked for 2 years as a biology teacher in an international school. Between 1992 and 2002, he worked as an expert in fisheries and aquaculture in various countries. He holds a Master's degree in Agricultural Sciences from Wageningen University in the Netherlands.

## Griet Van den Eynde

Griet Van den Eynde is the coördinator of the test development team at the Secondary Education Examination Board in Brussels. She manages the team of testing experts and subject matter experts who are responsible for every aspect of the testing process. She joined the Examination Board in 2016 after a 18-year career as a language, communication and PR teacher. Griet earned a master's degree and teaching qualification in Germanic philology from KU Leuven in Belgium.

## Soetkin De Knijf

Soetkin De Knijf began her teaching career in 2002 as a history teacher in secondary school. She joined the Secondary Education Examination Board in 2015 as a specialist of history exams. In 2018, she made the move to the Qualifications and Curriculum department, where she works as a policy advisor.

## Dries Vrijsen

Dries Vrijsen is the Process Manager for Analysis and Subject Coordinator for Mathematics at the Secondary Education Examination Board in Brussels. In this role, he oversees math exams and guides the quality control process for both exams and exam questions. He joined the Examination Board in 2019 after a 16-year career as a mathematics teacher at WICO Pelt secondary school. Dries earned a master's degree and teaching qualification in mathematics from KU Leuven in Belgium.

# HIGH-QUALITY SUMMATIVE E-TESTING AT THE EXAMINATION BOARD FOR SECONDARY EDUCATION IN FLANDERS

## Abstract

The Flemish Examination Board for Secondary Education gives everyone, regardless of age, nationality, or school background, the opportunity to obtain a diploma in secondary education through self-study. The examination programme that participants must complete consists mainly of digital exams to be taken at the professionally equipped examination centre in Brussels. More than 45,000 exams are conducted on an annual basis.

In 2012, the Examination Board started the transition from written exams to digital exams and has since been permanently committed to improvements through technological developments. E-testing has made the process of conducting exams more efficient. They preferably consist of close-ended questions, which enables automatic correction. Developing these type of questions requires specific expertise and adequate staff training, which are crucial for e-testing.

Test items are created based on a set of test principles (validity, reliability, user-friendliness, and authenticity). The test items are brought together in exams that are composed based on test matrices to ensure equivalence between exam versions.

Quality control is based on psychometric analysis. All digital exams are automatically graded. The analysis aims to measure, monitor, and improve the quality of our items and exams. It allows the test developers to optimise and critically review each test item.

Summative e-testing procedures are written down in this paper. The different phases of the testing cycle, from selecting learning objectives and construction of the test to correction and analysis, are illustrated using examples from science exams. The paper ends with a critical review of e-testing at the Flemish Examination Board and a look at future developments.

## Introduction

The Examencommissie secundair onderwijs, as the Secondary Education Examination Board is officially called in Flanders, offers everyone an alternative to obtaining a certificate or diploma of secondary education through self-study and thus without going to school. Currently, almost 10,000 participants are enrolled, which is about 2% of the total student population in Flemish secondary education. They can choose between a limited number of education programmes that lead to a Flemish certificate or diploma in general or vocational education. The examination programmes test all final learning objectives in secondary education set by the Government of Flanders. Depending on the degree and level of education, each examination programme consists of 10 to 15 summative exams.

***Full-time secondary education in Flanders*** *is subdivided into three stages of two grades each*[1]*. Secondary education aims to provide young people with the necessary competencies for personal development, social participation, further education, or a profession. With regard to the structure of secondary education, a distinction is made in the second and third stages between education types and orientations. Each programme in these stages is defined by these features.*

*Orientations relate to the paths which pupils can choose after finalising a programme in secondary education. There are* ***three different orientations****:*

- Orientation to higher education prepares pupils for a smooth transition to higher education
- Both a vocational and orientation to higher education prepares pupils for a smooth transition to higher education or the labour market
- Vocational orientation prepares pupils for a smooth transition to the labour market

*Starting in the 2023-2024 academic year, all primary and secondary schools will administer* ***central exams*** *to support the internal quality of schools and strengthen the quality of education. The exams, which are adaptive and low stakes, initially focus on Dutch and mathematics and will be administered and processed digitally. In 2024, exams will be administered in the 4th year of primary education and the 2nd year of secondary education. In 2026, the exams will also be rolled out in the last year of primary education, and in the last year of secondary education in 2027. They provide additional information (next to their own evaluation) that a school can use as a basis for deliberation.*

Most exams are digital but there are also oral exams for language subjects and practical exams in which professional competencies are tested in vocational education.

Participants can register and start planning their exams at any time. There are no admission requirements. Most participants are between the ages of 15 and 19. They are homeschooled or have left school partially or completely, for example, because of school fatigue, mental or physical problems, an intensive sports career, a long-term stay abroad, or imprisonment. Each participant has their own unique reason for participating in the examination board programme.

As an examination board, we do not provide education or guidance but expect participants to prepare by themselves or to hire a private tutor. We do, however, publish info documents, called *vakfiches*, describing all the goals and skills a participant must master to successfully sit an exam.

We organise five series of exams per year in which each subject of the education programme appears. If a participant fails an exam, they may sit for it again. Each subject can be taken up to three times a year. In 2023, the Examination Board organised more than 45,000 individual exams.

Until 2012, we organised only oral or written exams. From then on, we started digitising the exams to work more efficiently and qualitatively.

[1] The matrix of secondary education in Flanders (2023). https://www.youtube.com/watch?v=8036nc2mEK4

## An integrated assessment policy for e-testing

To meet high-quality e-testing standards necessary for summative exams, we use an integrated and circular assessment policy as illustrated by the testing cycle in Figure 1. Since it is an ongoing and never-ending process, each component of the cycle influences the other components. It is crucial to emphasise that qualitative testing encompasses the entire testing cycle. With the necessary quality checks in place, it facilitates a gradual enhancement in the quality of our exams

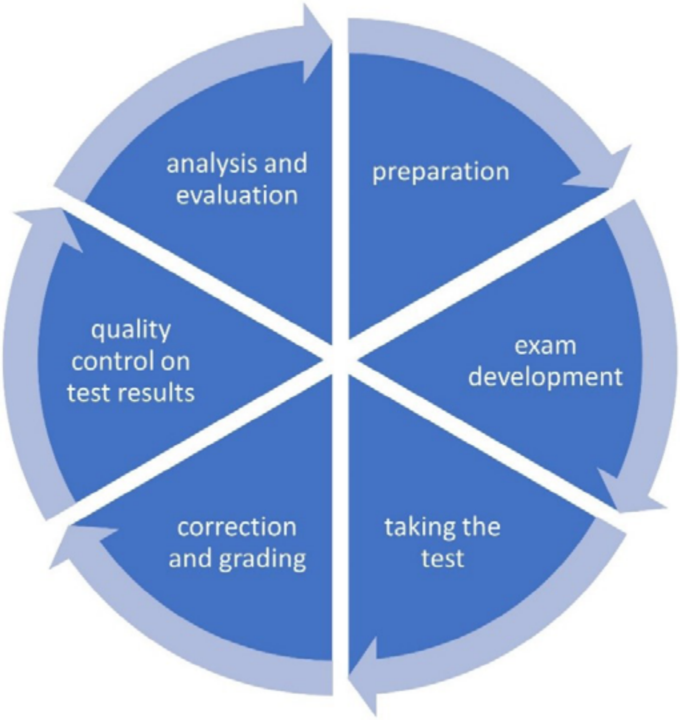


Figure 1: An integrated assessment policy.

In this paper, we briefly discuss the six components of the testing cycle, focusing mainly on exam development, grading, and quality control.

## Component 1 Preparation

In the preparation phase, we develop the exam programme for each education programme and create ‘vakfiches’: info documents for each exam intended for our participants. These documents include the learning objectives and provide information on exam procedures, grading, and suggestions for course materials [2]. The learning objectives are derived from the final learning objectives set by the Government of Flanders [3]. Before publication, these info documents undergo validation by the Flemish Inspectorate of Education [4].

[2] Examencommissie secundair onderwijs, 2024. https://examencommissiesecundaironderwijs.be/

[3] Onderwijsdoelen Vlaanderen, 2024. https://onderwijsdoelen.be/ and Learning objectives for secondary education in Flanders (2023). https://www.youtube.com/watch?v=HnmYMvYyHkg

[4] Vlaamse onderwijsinspectie, 2024. https://www.onderwijsinspectie.be/

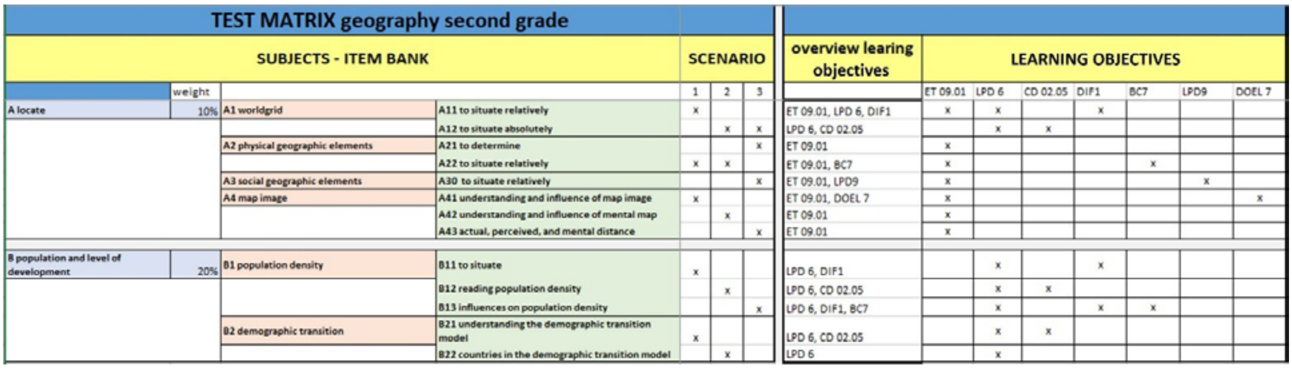
# Component 2 Exam development

The starting point of exam development is the design of a test matrix. After that, the development of test items can begin.

## The importance of a test matrix

As soon as the info documents have been approved, we use them to construct test matrices (Figure 2). The structure of the matrix reflects the learning objectives, organised into components that are also represented in our exams. Each component has a fixed weight and the number of questions within a component is proportional to its weight. On average, an exam consists of 40 items. The test matrix ensures content validity and that all final learning objectives are tested in a sufficiently representative manner.

Different scenarios are designed to ensure equivalence between exam versions. These scenarios outline the different test items within each learning objective. Based on analyses conducted after the exams are taken, average scores determine whether the different scenarios indeed generate equivalent exams. Further details are provided in the subsequent section on analysis.



| TEST MATRIX geography second grade | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SUBJECTS - ITEM BANK | | | | SCENARIO | | | overview learing objectives | LEARNING OBJECTIVES | | | | | | |
| | weight | | | 1 | 2 | 3 | | ET 09.01 | LPD 6 | CD 02.05 | DIF1 | BC7 | LPD9 | DOEL 7 |
| A locate | 10% | A1 worldgrid | A11 to situate relatively | x | | | ET 09.01, LPD 6, DIF1 | x | x | | x | | | |
| | | | A12 to situate absolutely | | x | x | LPD 6, CD 02.05 | | x | x | | | | |
| | | A2 physical geographic elements | A21 to determine | | | x | ET 09.01 | x | | | | | | |
| | | | A22 to situate relatively | x | x | | ET 09.01, BC7 | x | | | | x | | |
| | | A3 social geographic elements | A30 to situate relatively | | | x | ET 09.01, LPD9 | x | | | | | x | |
| | | A4 map image | A41 understanding and influence of map image | x | | | ET 09.01, DOEL 7 | x | | | | | | x |
| | | | A42 understanding and influence of mental map | | x | | ET 09.01 | x | | | | | | |
| | | | A43 actual, perceived, and mental distance | | | x | ET 09.01 | x | | | | | | |
| B population and level of development | 20% | B1 population density | B11 to situate | x | | | LPD 6, DIF1 | | x | | x | | | |
| | | | B12 reading population density | | x | | LPD 6, CD 02.05 | | x | x | | | | |
| | | | B13 influences on population density | | | x | LPD 6, DIF1, BC7 | | x | | x | x | | |
| | | B2 demographic transition | B21 understanding the demographic transition model | x | | | LPD 6, CD 02.05 | | x | x | | | | |
| | | | B22 countries in the demographic transition model | | x | | LPD 6 | | x | | | | | |

Figure 2: The general layout of a test matrix.

Based on the test matrix, the item bank architecture is established in our digital testing platform, assessmentQ. The lowest subject levels contain the items. We employ automated scenarios to select items. Exams are automatically generated from the item bank and the scenarios we have created.

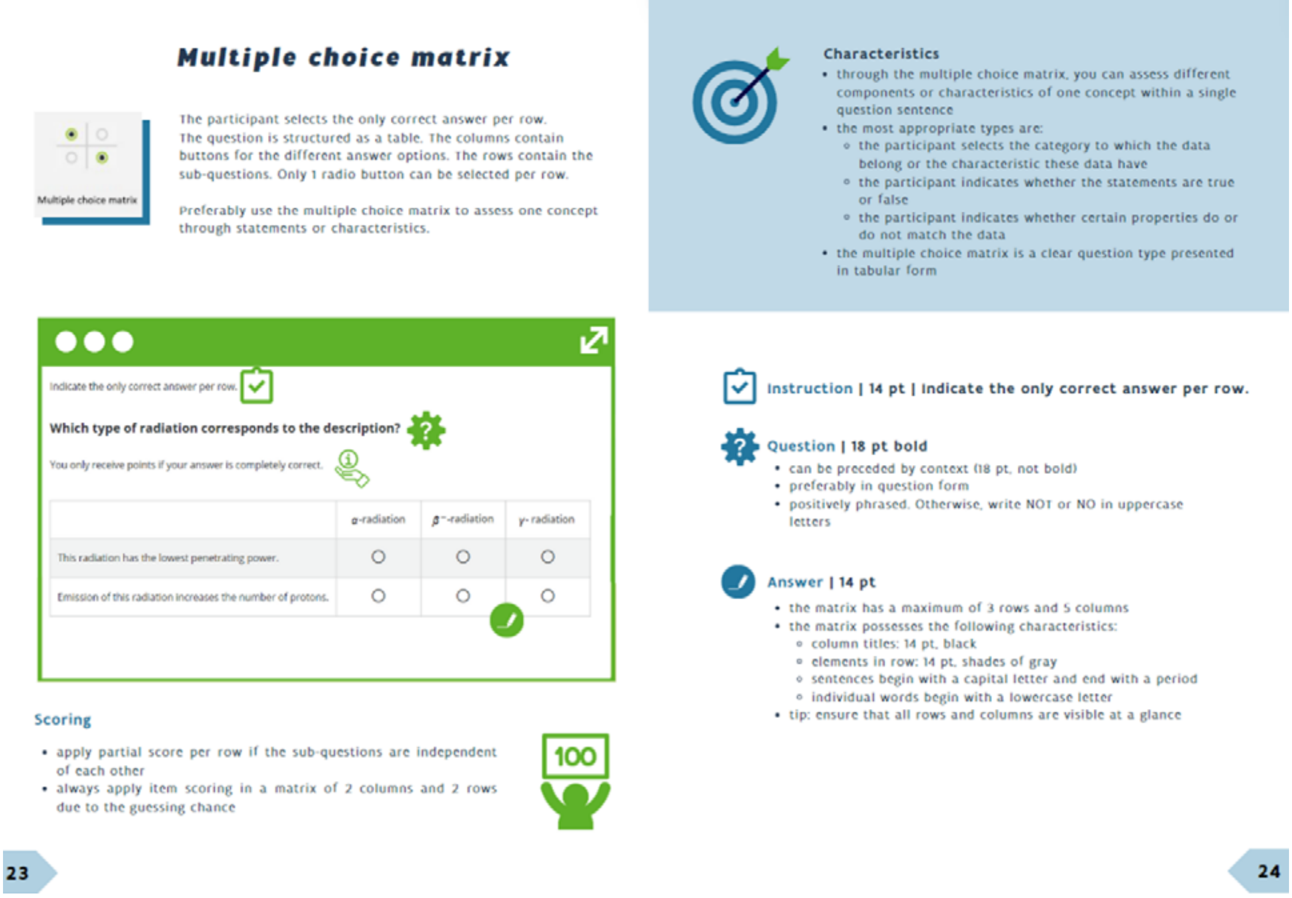
Below are the fundamental principles of our exam development process:

- tests are 100% digital unless they involve speaking skills or practical competencies;
- all tests are summative;
- test items are slim. They have a single-question format and include only content strictly necessary to answer the questions;
- close-ended questions are preferred. They allow for fast, objective, and automated correction. Open-ended questions are only allowed to meet validity standards.

## How to construct a test item?

After designing a test matrix, test items are developed according to standard procedures. These are then included in a database that forms the basis for the exams.

We have our own manual for item construction. This booklet delineates the test principles, including tips and tricks for choosing the appropriate item types and creating reliable test items (Figure 3). We use a variety of item types.



**Multiple choice matrix**

The participant selects the only correct answer per row. The question is structured as a table. The columns contain buttons for the different answer options. The rows contain the sub-questions. Only 1 radio button can be selected per row.

Preferably use the multiple choice matrix to assess one concept through statements or characteristics.

Indicate the only correct answer per row.

**Which type of radiation corresponds to the description?**

You only receive points if your answer is completely correct.

| | $\alpha$-radiation | $\beta^-$-radiation | $\gamma$- radiation |
|---|---|---|---|
| This radiation has the lowest penetrating power. | ○ | ○ | ○ |
| Emission of this radiation increases the number of protons. | ○ | ○ | ○ |

**Scoring**

- apply partial score per row if the sub-questions are independent of each other
- always apply item scoring in a matrix of 2 columns and 2 rows due to the guessing chance

**Characteristics**

- through the multiple choice matrix, you can assess different components or characteristics of one concept within a single question sentence
- the most appropriate types are:
  - the participant selects the category to which the data belong or the characteristic these data have
  - the participant indicates whether the statements are true or false
  - the participant indicates whether certain properties do or do not match the data
- the multiple choice matrix is a clear question type presented in tabular form

**Instruction | 14 pt | Indicate the only correct answer per row.**

**Question | 18 pt bold**

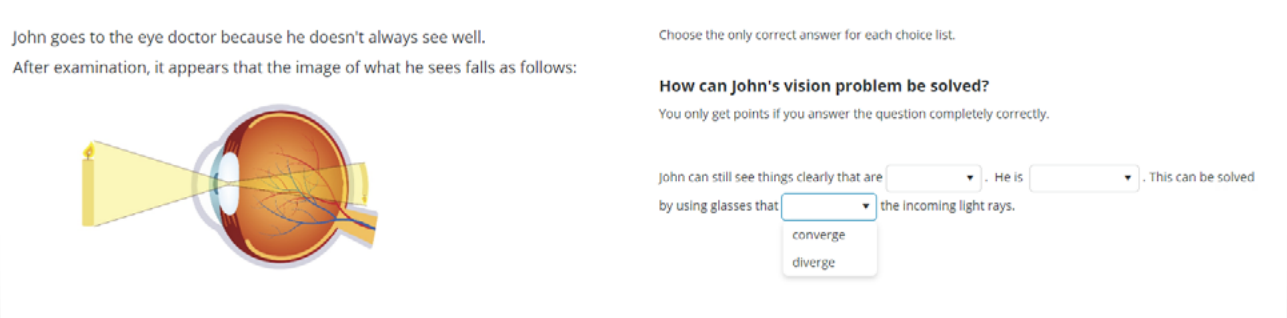
- can be preceded by context (18 pt, not bold)
- preferably in question form
- positively phrased. Otherwise, write NOT or NO in uppercase letters

**Answer | 14 pt**

- the matrix has a maximum of 3 rows and 5 columns
- the matrix possesses the following characteristics:
  - column titles: 14 pt, black
  - elements in row: 14 pt, shades of gray
  - sentences begin with a capital letter and end with a period
  - individual words begin with a lowercase letter
- tip: ensure that all rows and columns are visible at a glance

Figure 3: A double page of the guide for test development stipulates the instructions for the multiple-choice matrix item type.

The test developer's first task is to determine what goal they want to achieve with the test item. Depending on this goal, they select the most suitable item type. A drop-down item type, for example, is very suitable for testing reasoning ability as illustrated in Figure 4.



John goes to the eye doctor because he doesn't always see well.

After examination, it appears that the image of what he sees falls as follows:

Choose the only correct answer for each choice list.

**How can John's vision problem be solved?**

You only get points if you answer the question completely correctly.

John can still see things clearly that are [ ▾ ] . He is [ ▾ ] . This can be solved by using glasses that [ ▾ ] the incoming light rays.

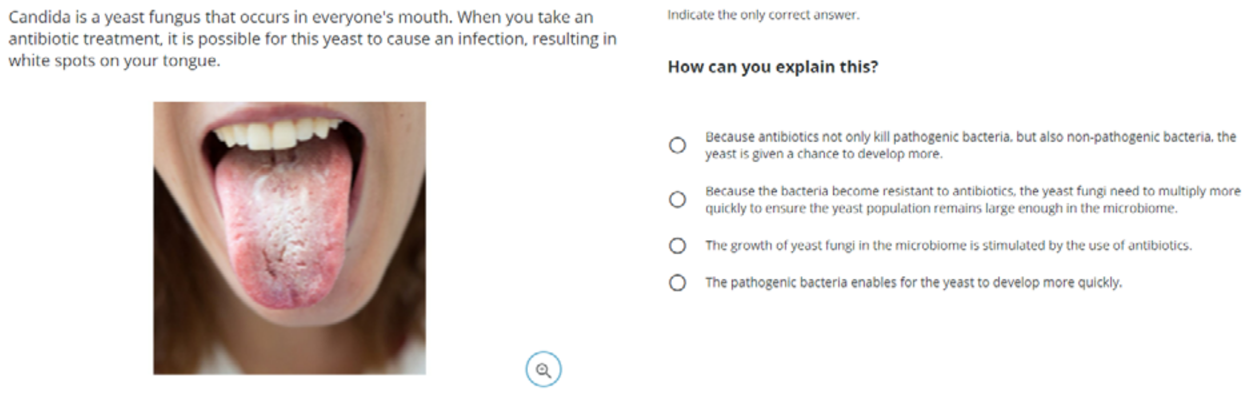
Figure 4: An example of a drop-down item type from a biology exam.

When constructing test items, we make sure to stick to four testing principles: validity, reliability, transparency or user-friendliness, and authenticity. These principles are illustrated in the following sections through examples of test items drawn from various science exams.

## How to guarantee item validity?

Item validity refers to the extent to which a test item accurately measures what it is intended to measure. A valid item assesses the knowledge or skills according to the final learning objectives translated into the info documents.

An example of a valid item from a biology exam for the second grade of Flemish secondary education is shown in Figure 5. The learning objective states, 'You explain that diseases result from a disruption of the balance between organisms'. With this question, the participants must demonstrate their understanding of how antibiotics work and what their impact is on disrupting the bacterial balance.



Candida is a yeast fungus that occurs in everyone's mouth. When you take an antibiotic treatment, it is possible for this yeast to cause an infection, resulting in white spots on your tongue.

Indicate the only correct answer.

**How can you explain this?**

○ Because antibiotics not only kill pathogenic bacteria, but also non-pathogenic bacteria, the yeast is given a chance to develop more.

○ Because the bacteria become resistant to antibiotics, the yeast fungi need to multiply more quickly to ensure the yeast population remains large enough in the microbiome.

○ The growth of yeast fungi in the microbiome is stimulated by the use of antibiotics.

○ The pathogenic bacteria enables for the yeast to develop more quickly.

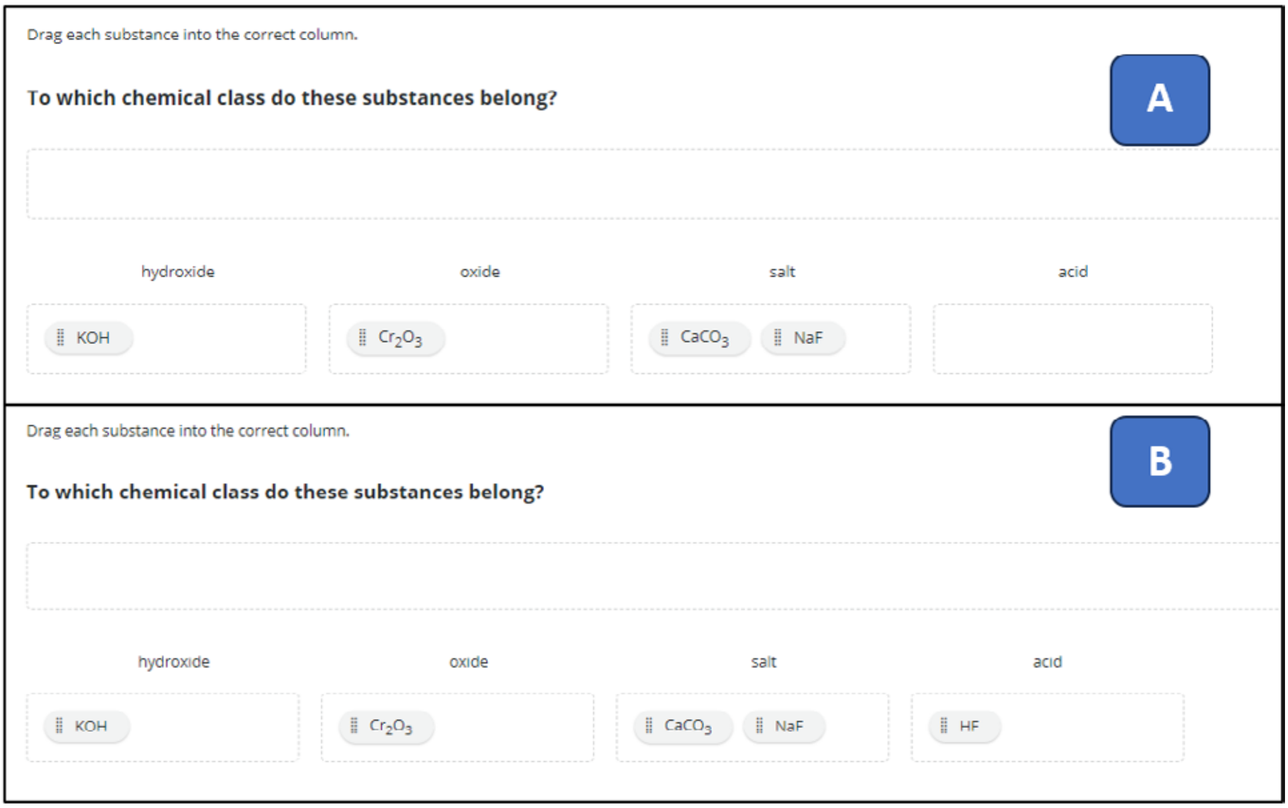
Figure 5: Example of an item from a biology exam to illustrate the principle of validity.

## How is item reliability guaranteed?

Reliability of test items refers to the consistency and stability of the scores obtained from the items over repeated administrations. A reliable test item consistently yields similar results when administered to a similar group under similar conditions.

This can be accomplished through various analyses after the test is taken (see the section on analysis). Furthermore, employing clear and unambiguous language in the test item and providing adequate instructions contribute to enhancing reliability. Therefore, we do not use unnecessary complex phrasing or wording and avoid misleading or tricky questions. Scoring must be done equally and fair. There should be a consensus on the correct answer and the distractors.

To illustrate this, two versions of almost the same item from a chemistry exam are presented in Figure 6. For each question, it should be clear to participants what the expected answer is, assuming they have the necessary knowledge. In version A, four answers are to be placed in four columns, while one column remains empty. In version B, five answers should be allocated to four columns. Version A can be perceived as less reliable and potentially misleading, as a student might anticipate distributing four items evenly across the columns, which is not the correct answer.



Drag each substance into the correct column.

To which chemical class do these substances belong?

| hydroxide | oxide | salt | acid |
|---|---|---|---|
| KOH | $Cr_2O_3$ | $CaCO_3$<br>NaF | |

Drag each substance into the correct column.

To which chemical class do these substances belong?

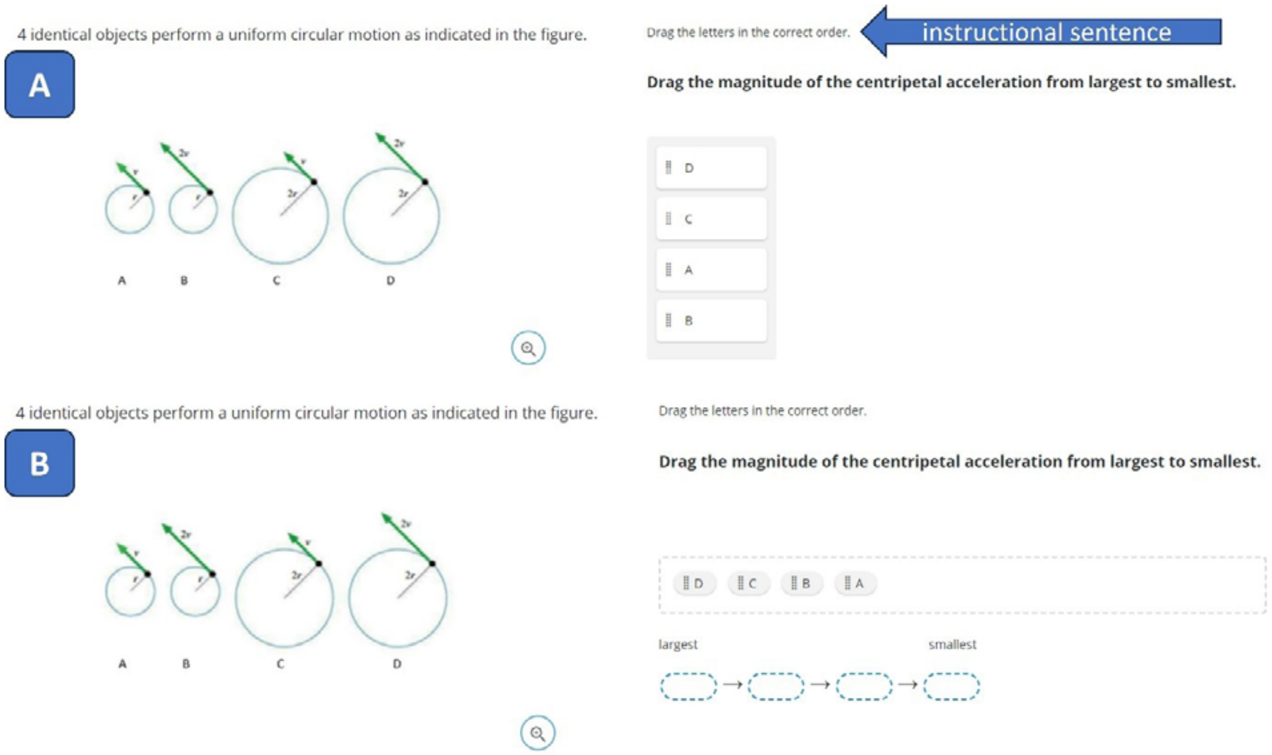
| hydroxide | oxide | salt | acid |
|---|---|---|---|
| KOH | $Cr_2O_3$ | $CaCO_3$<br>NaF | HF |

Figure 6: Example of an item from a chemistry exam to illustrate the principle of reliability.

## How is transparency or user-friendliness guaranteed?

Transparency or user-friendliness refers to the clarity and accessibility of the assessment process for participants. These procedures enable students to fully understand the expectations and requirements of the assessment, accommodating diverse learners, and ensuring that all participants can effectively demonstrate their abilities regardless of background or learning style. Therefore, we implement a clear and standardised layout, provide clear instructions, and establish understandable scoring rules. We select item types based on the learning objectives while considering user-friendliness as illustrated in the example in Figure 7.

Two versions of nearly identical items from a physics exam are presented in Figure 7. Each item should clearly communicate to participants what the expected answer is, assuming they possess the requisite knowledge: a context is provided, followed by an instructional sentence, and a clear question. In version A, we observed that dragging letters to another location causes issues, as they may inadvertently swap places. Reconstructing the previous item as shown in version B appears to enhance user-friendliness without altering the content.



A

4 identical objects perform a uniform circular motion as indicated in the figure.

Drag the letters in the correct order.

instructional sentence

Drag the magnitude of the centripetal acceleration from largest to smallest.

D
C
A
B

B

4 identical objects perform a uniform circular motion as indicated in the figure.

Drag the letters in the correct order.

Drag the magnitude of the centripetal acceleration from largest to smallest.

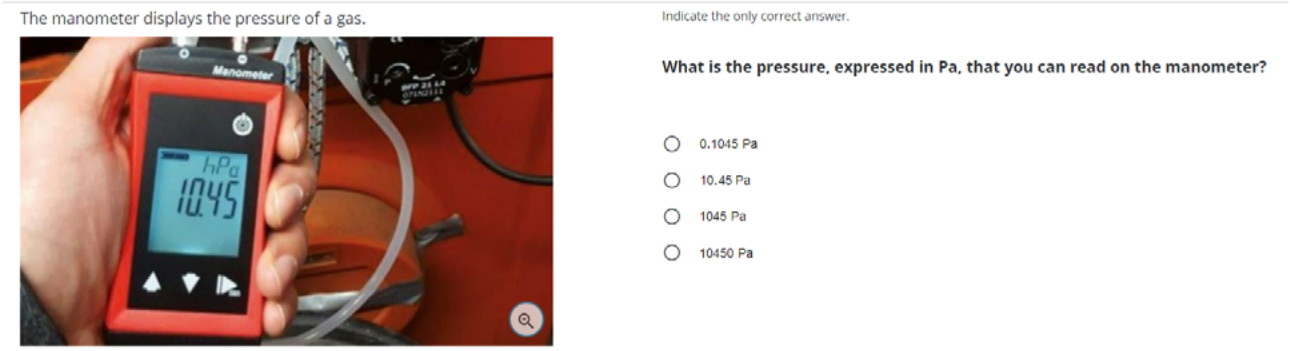
D C B A

largest → → → smallest

Figure 7: Example of an item from a physics exam to illustrate the principle of user-friendliness.

## How is authenticity guaranteed?

Authenticity refers to the degree to which the test reflects real-world situations or contexts that are relevant to the learners' experiences. Authentic assessments often involve real-world problem-solving, critical thinking, and the application of knowledge. Therefore, a lifelike context is used as in the example in Figure 8. The apparatus depicted on a natural science exam represents a digital manometer and the provided values are realistic.



The manometer displays the pressure of a gas.

Indicate the only correct answer.

What is the pressure, expressed in Pa, that you can read on the manometer?

○ 0.1045 Pa

○ 10.45 Pa

○ 1045 Pa

○ 10450 Pa

Figure 8: Example of an item from a physics exam to illustrate the principle of authenticity.

## Continuous screening of test items

Each test item undergoes screening by at least one experienced colleague in the test development section before its inclusion in the item bank. Items are regularly reviewed in group discussions with colleagues from diverse subject backgrounds to facilitate mutual learning and enhance the quality of the items. Additionally, there is a permanent working group consisting of ten members of our team who address issues related to drafting good test items, evaluating new item types, and refining and validating guidelines for item drafting. Item creation is an ongoing process where we continuously improve our item banks through screenings and psychometric analyses.

## Component 3 Taking the test

After the test items are developed and compiled into an exam, participants take the test in our well-equipped exam centre (Figure 9). Participants must have optimal conditions during the test to ensure that the reliability of the test isn't compromised in this phase of the test cycle. That's why we chose to implement the principles of Universal Design for Assessment (UDA). The test environment must ensure that everyone, regardless of their age, size, ability, or disability, can take the test under optimal conditions.

To ensure reliability, we are strongly committed to fraud prevention and detection with well-trained supervisors. We have strict guidelines on storing personal items such as mobile phones and smartwatches, and the covering of exam computers. The information participants are allowed to look up on the internet during the exam, is exclusively accessible on whitelisted websites.

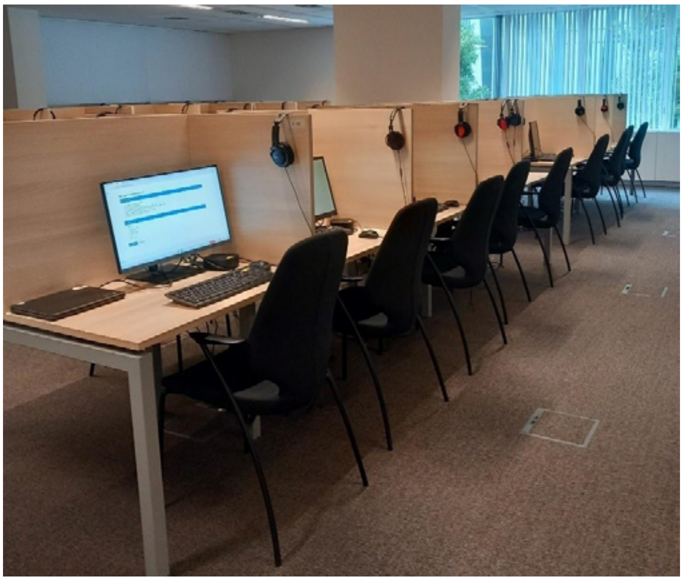


Figure 9: The exam centre of the Examination Board in Brussels.

## Component 4 Correction and grading

The advantage of exams with closed items is that the correction is done automatically. The results are immediately available for quality control, and there are no discrepancies in assessment because there is no human intervention.

Exams with open-ended questions require human correction, which is time-consuming and poses a risk of variations in assessment. Regardless of how stringent the correction regulations are, achieving consensus among raters is very challenging, especially when working with a broad network of external raters.

## Component 5 Quality control and test results

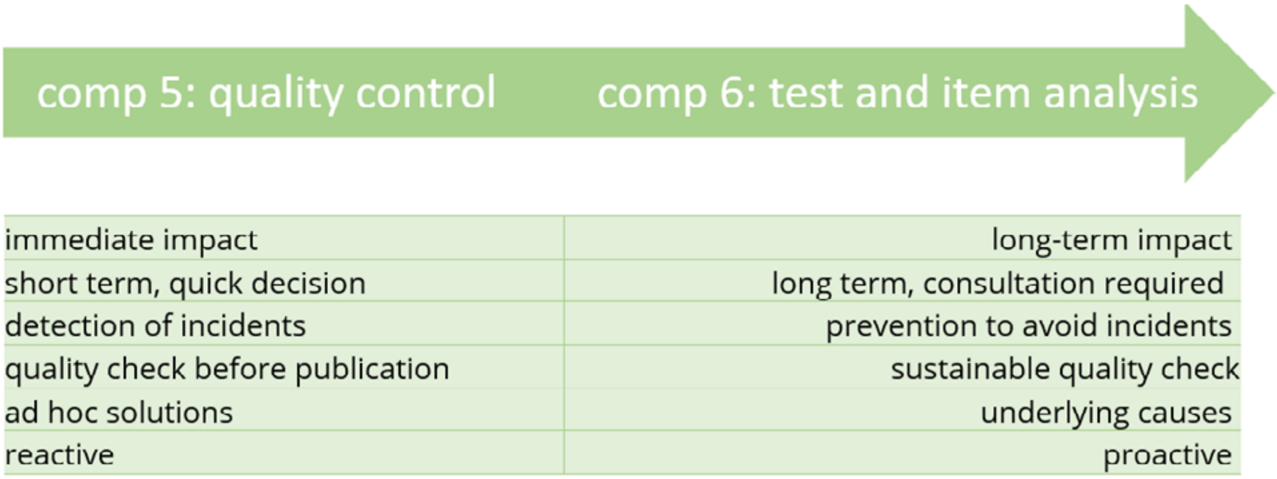
Before we publish test results, we perform an initial quality control to avoid errors. According to our exam regulations, the grades can only be modified in favour of the participant in order to avoid damage to our public image and the confidence in our organisation.

This phase of quality control begins as soon as the results of the exams are visible and the correction work for possible open-ended questions is delivered. This period lasts a maximum of seven days, during which we have time to verify if anything went wrong with the test items, the test-taking, or the correction.

Quality control (component 5) and test and item analysis (TIA) (component 6) are strongly linked. Both evaluations are performed to detect anomalies during the quality check but also for long-term improvement of our item banks, as shown in Figure 10.

The main focus of quality control is to correct errors due to poor item or test development that impact our participants' final scores. If needed, we take action to erase this impact. This can be done, for instance, by adjusting scores or eliminating items. However, the seven day period is too short to overthink the entire process of item and test development. During the analysis, we delve deeper and examine poorly and well-performing items to improve the items in our database.

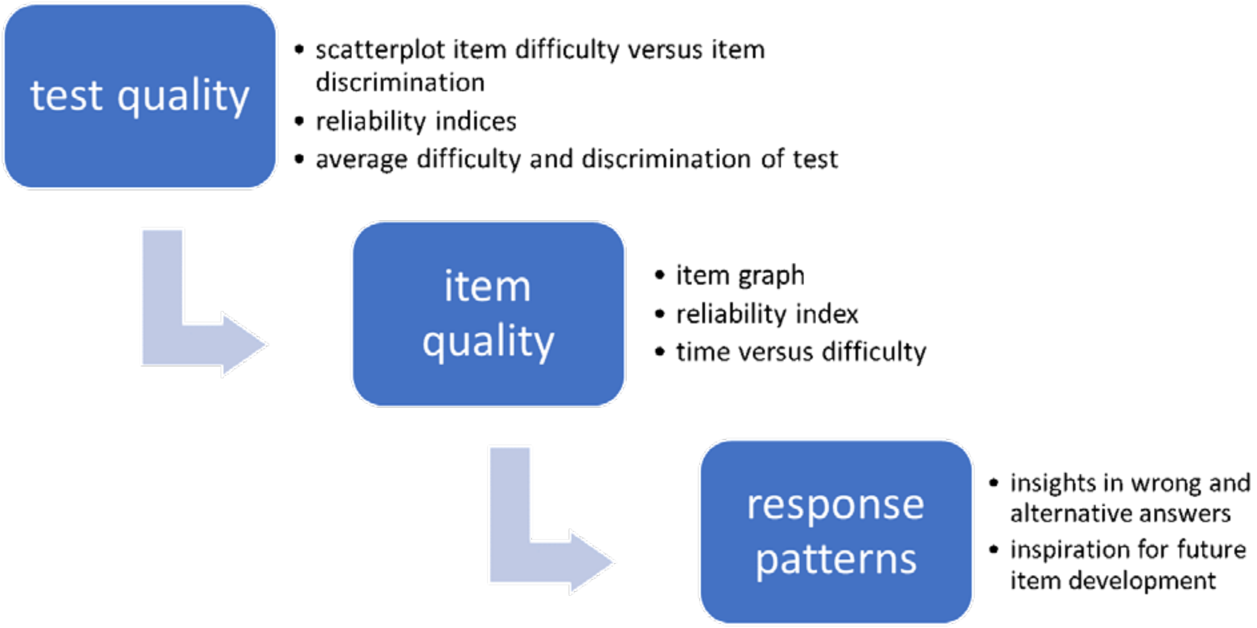
## Component 6 Analysis and evaluation



| | |
|---|---|
| immediate impact | long-term impact |
| short term, quick decision | long term, consultation required |
| detection of incidents | prevention to avoid incidents |
| quality check before publication | sustainable quality check |
| ad hoc solutions | underlying causes |
| reactive | proactive |

Figure 10: Quality control versus test and item analysis.

Whether it is a quality control or an analysis, the measures and tools we use are more or less the same. We summarise the difficulty, discrimination, and reliability of the test and items in different measures, while response patterns give qualitative data about the item. The analysis performance scheme is shown in Figure 11.



test quality

- scatterplot item difficulty versus item discrimination
- reliability indices
- average difficulty and discrimination of test

item quality

- item graph
- reliability index
- time versus difficulty

response patterns

- insights in wrong and alternative answers
- inspiration for future item development

Figure 11: Analysis performance scheme.

Lastly, we conduct specific investigations that fall outside the scope of this article. These include, for example, an inter-rater reliability study or an investigation into differences between subgroups.

The most important part of this component consists of performing an analysis of the test results and the quality of the items in the tests to monitor the general quality of our evaluation.

## Test quality

To get a general view of the difficulty level of the exam, we measure item difficulty, the P-value of an item, and its discriminatory power or RIT-value. P-value and RIT-value are parameters that vary from 0 to 100%. P=100 means that every participant answered the item correctly, P=0 means that no one gave the correct answer to this question. The RIT-value shows the correlation between the item score and the test score. We consider items with RIT < 20 to have low discriminatory power. It means that participants with high overall scores on the test are not consistently performing better on these items compared to participants with lower overall scores.

If we plot the P-value of all items on the x-axis and the RIT-value on the y-axis, we get a scatterplot. As can be seen in Figure 12, we use thresholds to categorise items. In the upper left, items are difficult but discriminate between high-scoring and low-scoring participants. The graph shows a nice overall spread. Items in the lower left should be investigated, they are assessed as difficult, yet they have poor correlation to the final test score.

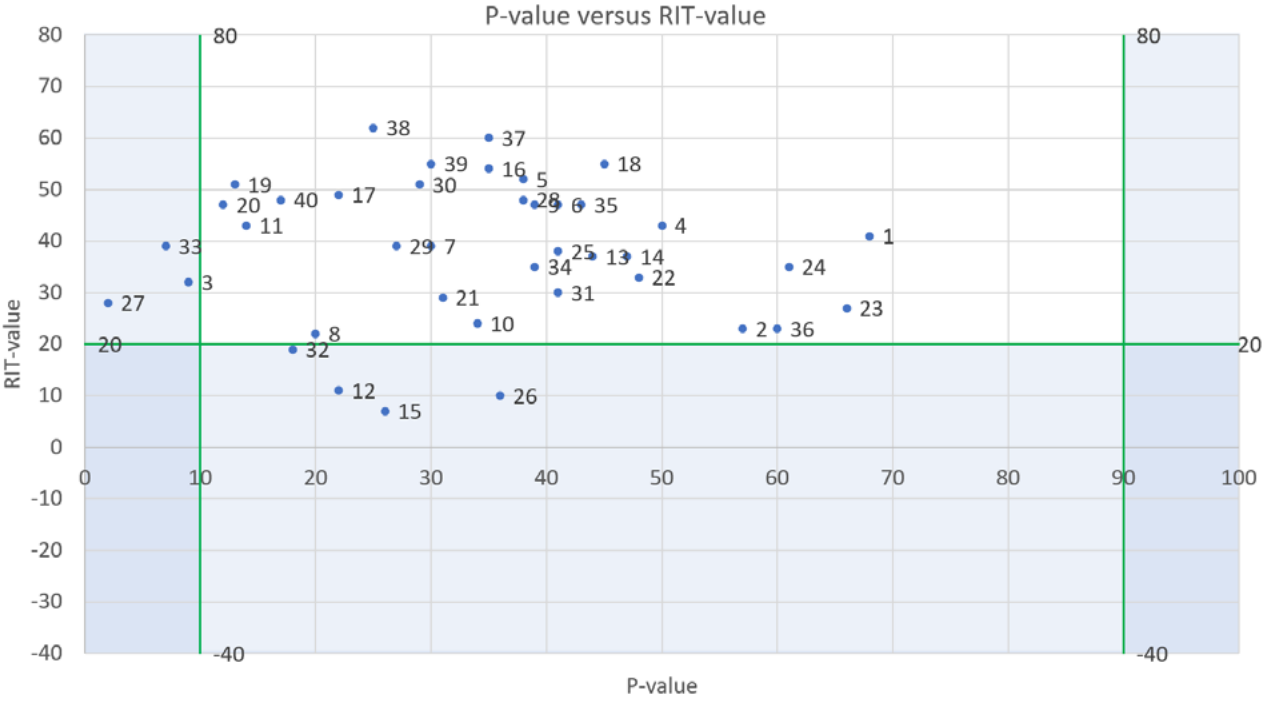


Figure 12: Scatterplot of P- and RIT-values.

All these measures can be summarised in the average P-value and average RIT-value of the test. We keep track of these values over time to identify different behaviours of the scenarios and to monitor the overall difficulty. If a peak in the graph corresponds to a particular scenario, we probably have to adjust it and get it in line with the rest. Figure 13 shows the average P-value of 48 consecutive tests generated by 7 randomly chosen scenarios.

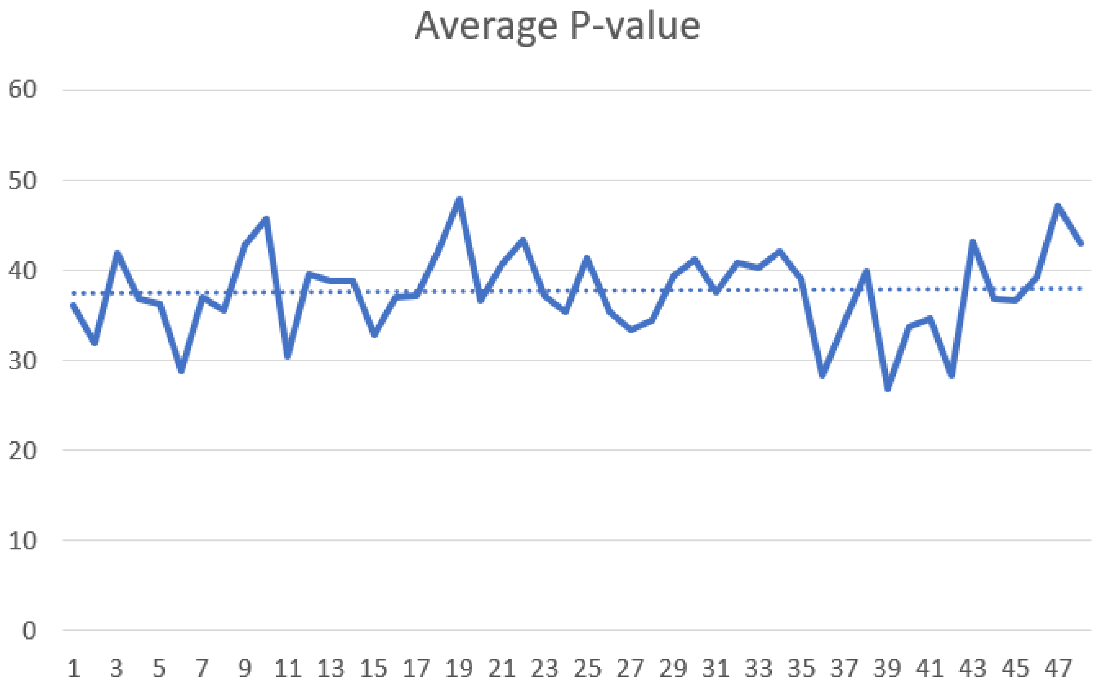


Figure 13: Evolution of test difficulty over time.

To measure the overall reliability of our exams, we use two parameters: Cronbach's Alpha and Guttman's Lambda-2 (Figure 14). They both measure internal consistency and can be calculated from a single test. These parameters should approach 1 in a reliable test. Once more, we use thresholds as guidance to interpret them, taking the shortcomings of these measures as described in the literature into account.

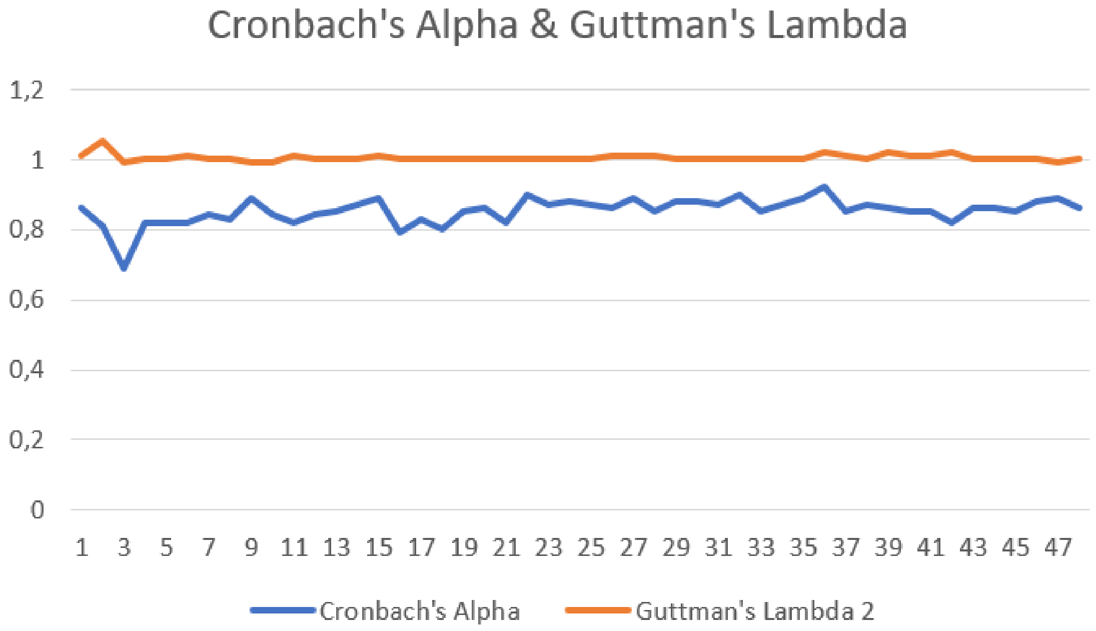


Figure 11 Evolution of Cronbach's Alpha and Guttman's Lambda-2.

This is how we monitor equivalence between different tests through balanced item banks and scenarios.

## Item quality

We attempt to identify both poorly and well-performing items through various measures and graphs. The former provides insight into further development of the item bank, while the latter needs adjustment.

Once again, we examine the difficulty and discriminatory power of each item. We supplement the P and RIT-values with an item graph. Participants' scores are sorted from low to high and divided into four groups, with high achievers in group 4 and low achievers in group 1. Within each group, the P-value of the items is listed. We expect to see a rising graph: high achievers are expected to answer the item correctly. If that is not the case, the distractors may mislead our strongest participants. Divergent patterns might indicate poor item performance. A sample item graph is shown in Figure 15.

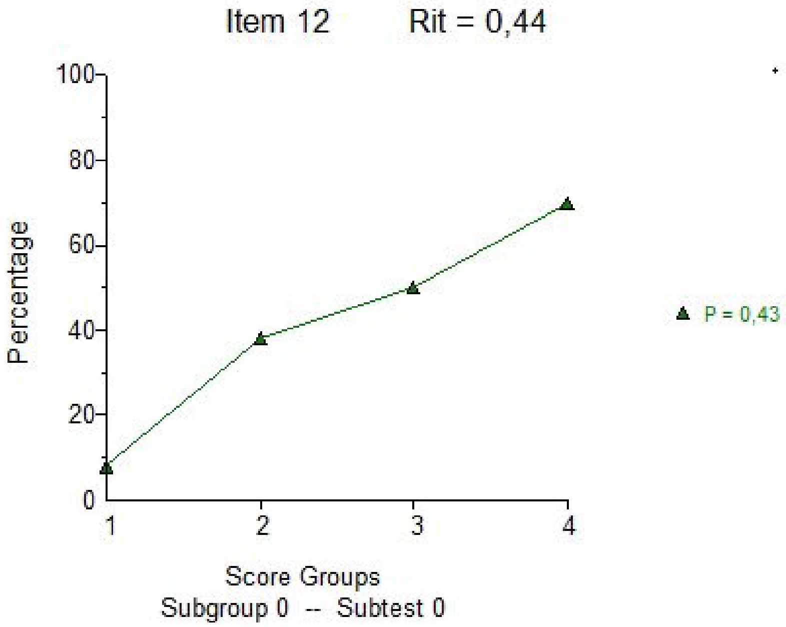


Figure 15 Item graph.

If we plot the average time it took participants to solve the item against the P-value of the item, we get an indication of transparency, as shown in Figure 16. Generally, we expect a declining trend line: difficult exercises usually require more time than easy exercises. Exercises with a high P-value (indicating easiness) and a long working time may suggest that the item is formulated in a challenging manner or is not concise enough. Items with a low P-value (indicating difficulty) and a short working time may suggest that they are skipped or incorrectly perceived in terms of difficulty by the candidate.

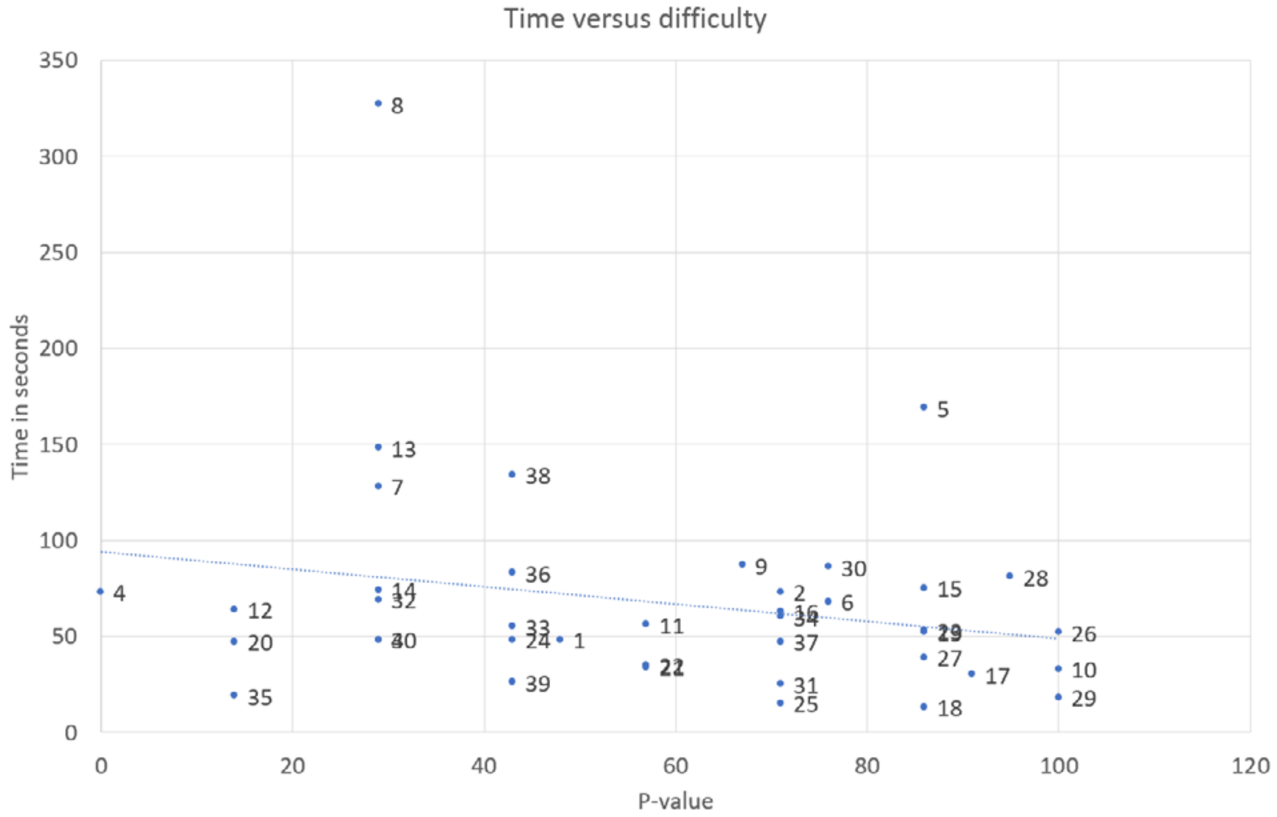


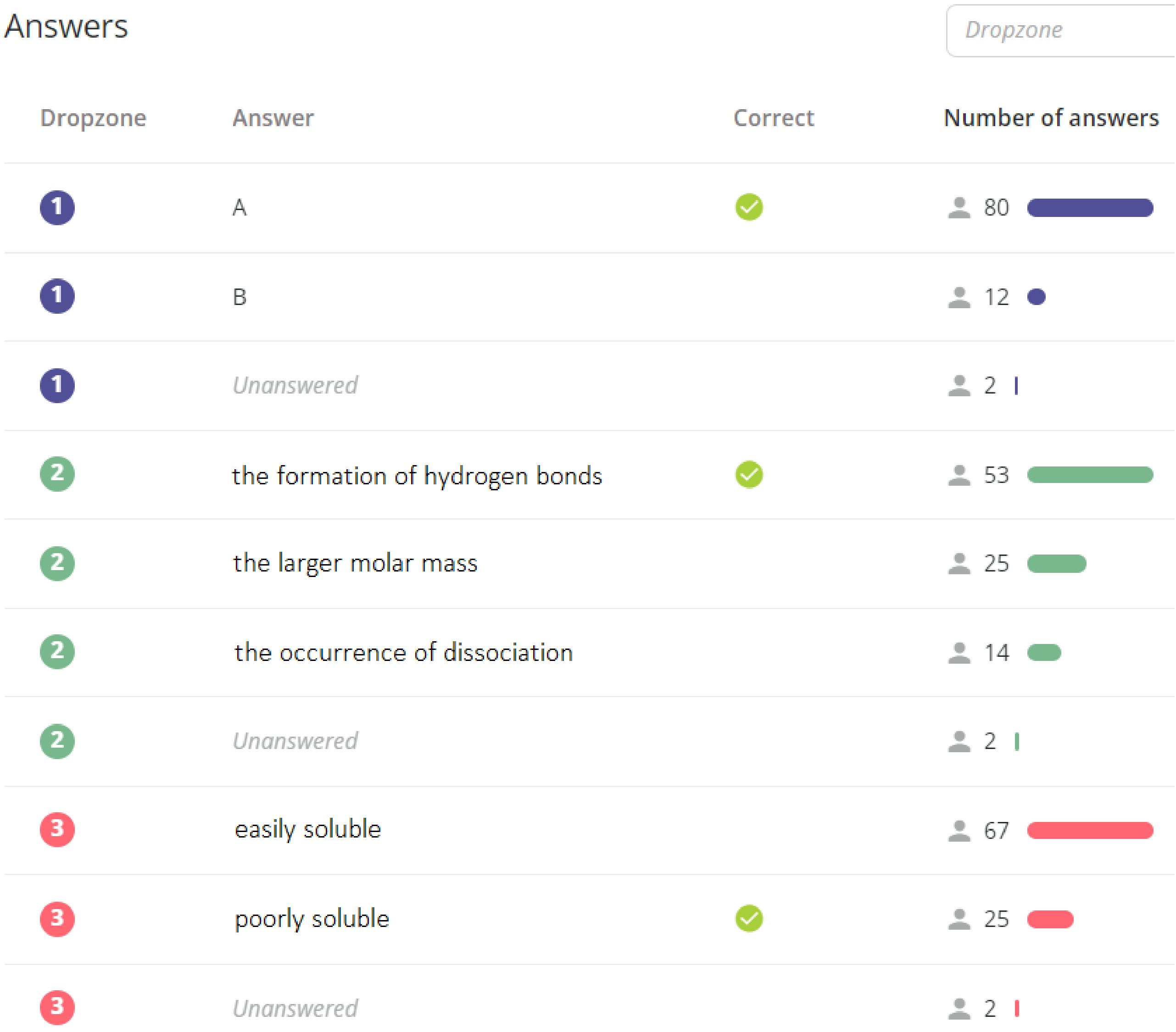
Figure 16 Time versus difficulty.

To get an idea of the contribution of an item to the overall internal consistency or reliability of the test, we compute the D-value, which is the difference between the reliability of the total test measured by the Alpha Cronbach and the Alpha Cronbach Remainder (AR) of the investigated item. The AR is the Alpha Cronbach of a test minus the item in question. The more positive the D-value, the greater the contribution of an item to internal consistency. An item with a D-value of 0 or with a negative D-value contributes negatively or does not contribute at all to reliability.

## Response patterns

In the next step, we investigate response patterns across different items. Quality control checks for other potentially correct answers, for example, in gap-fill questions or multiple-choice questions. In a long-term analysis, these patterns can be used in the future development of an item, for example, the top four answers in a gap-fill item can become good distractors for a multiple-choice item.

We obtain the distribution of the number of times the answer was given or selected from our exam platform, as illustrated for drop-down items in Figure 17.



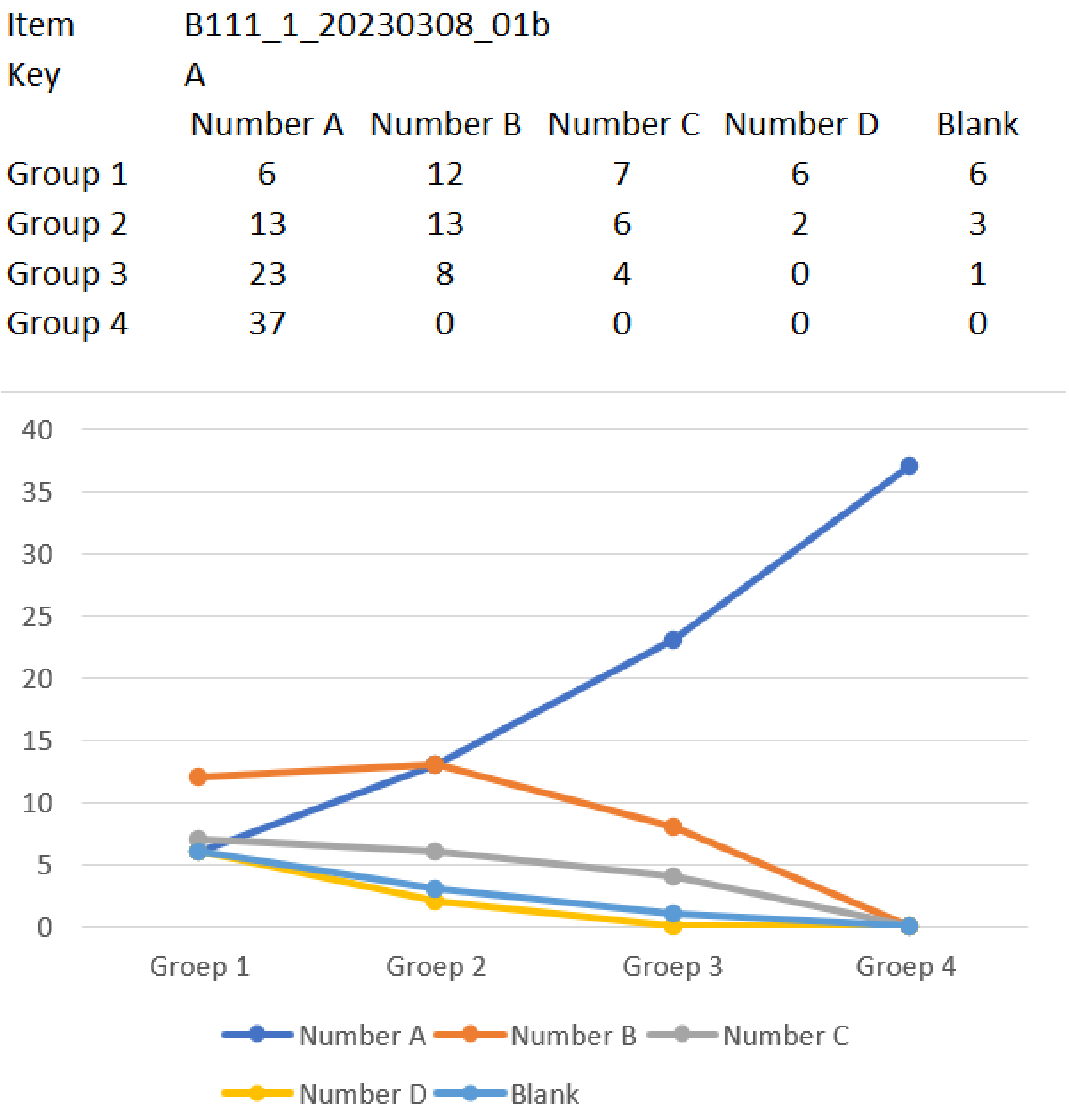
Answers

Dropzone

| Dropzone | Answer | Correct | Number of answers |
|---|---|---|---|
| 1 | A | ✓ | 80 |
| 1 | B | | 12 |
| 1 | *Unanswered* | | 2 |
| 2 | the formation of hydrogen bonds | ✓ | 53 |
| 2 | the larger molar mass | | 25 |
| 2 | the occurrence of dissociation | | 14 |
| 2 | *Unanswered* | | 2 |
| 3 | easily soluble | | 67 |
| 3 | poorly soluble | ✓ | 25 |
| 3 | *Unanswered* | | 2 |

Figure 17: Response patterns for drop-down items.

When examining multiple-choice items, one can also create an item graph for the distractors. This gives insight into which score groups are drawn to a specific distractor. Additionally, the discrimination between score groups by distractors becomes apparent. Figure 18 displays the item graph in blue, with distractors B, C, D, and blank shown in consecutive shades of orange, grey, yellow, and light blue. This item discriminates between score groups 1, 2, and 3, 4. Each distractor is chosen a fair number of times.



Item B111_1_20230308_01b

Key A

| | Number A | Number B | Number C | Number D | Blank |
|---|---|---|---|---|---|
| Group 1 | 6 | 12 | 7 | 6 | 6 |
| Group 2 | 13 | 13 | 6 | 2 | 3 |
| Group 3 | 23 | 8 | 4 | 0 | 1 |
| Group 4 | 37 | 0 | 0 | 0 | 0 |

Figure 18: Distractor graph.

We calculate the proportion of times each distractor is chosen. This results in an A-value for each distractor. The A-value for the correct answer is the same as the P-value for this item. A rule of thumb is that if 5% < A < P, the distractor is considered appropriate. If A < 5%, the distractor is unlikely to be chosen, and if A > P, the distractor might be too attractive.

Analysis results provide us insights into the performance of the tests and items. The scenarios and items in our databases are optimised based on the analysis results of the various exams. The analysis results serve as a starting point for discussions among colleagues. In these discussions, it is debated whether the items need to be adjusted, if there is an explanation for poorer analysis results, and suggestions are made for the optimisation and further development of exam questions.

Test and item analysis is never a standalone task; interpretation and processing of the results should play a major role. The methods we employ are part of classical test theory. Despite its known limitations, it is best suited for our approach to organising exams.

## Conclusion and reflection

The transition from traditional written exams to e-testing at the Flemish Examination Board for Secondary Education presents both advantages and challenges. The primary benefits of e-testing are its objectivity and efficiency in conducting exams, facilitated by the automatic grading of close-ended items. Additionally, the integration of technology allows for more exam flexibility; the inclusion of photos, audio, and film; the easy adaptation of items; and the sharing of item databases.

Moreover, the rigorous application of the principles of validity, reliability, transparency, and authenticity ensures assessment quality and fairness. The use of test matrices and scenario-based exams facilitates content validity and equivalence between exam versions. Continuous quality control mechanisms, including psychometric analysis, contribute to ongoing improvement in exam quality.

However, alongside these benefits, e-testing also has its challenges. Not all learning objectives can be thoroughly tested in a close-ended and digital context. Evaluating higher-order thinking skills and learning, social, and research competencies remains a challenge. Also, attention must be paid to ensuring the accessibility of the testing environment and fraud prevention.

Reflecting on our current state of e-testing, it is evident that while significant strides have been made, there is room for improvement and innovation. One promising avenue for future development is the integration of artificial intelligence (AI) into various aspects of the testing process. AI can be leveraged to enhance the correction of open-ended questions, which currently require human intervention and are prone to inconsistencies. AI-powered tools can provide more consistent and objective grading and reduce the time required for marking.

In conclusion, while the current e-testing framework at the Flemish Examination Board is robust and effective, embracing emerging technologies and continuous refinement of processes will be key to addressing existing challenges and enhancing the overall assessment experience. By integrating AI and other innovative solutions, the Board can ensure that its examination methods remain at the forefront of educational assessment, providing fair, reliable, and comprehensive evaluations for all learners.

## References

Albano (2018). Introduction to Educational and Psychological Measurement Using R, September 4, 2018.

Christopher, Desjardins, Okan Bulut (2018). Handbook of Educational Measurement and Psychometrics Using R. https://www.researchgate.net/publication/322173830

Cito (2013). TiaPlus User's Manual, M.& R. Department, https://tiaplus.cito.nl/TiaPlus_Users_Manual.pdf

Cito (2013). TiaPlus User's Manual. M. & R. Department, Cito, Arnhem, Netherlands.

Examencommissie secundair onderwijs (2024). https://examencommissiesecundaironderwijs.be/

Goldebeld P. (1992). Toets- en itemanalyse met TIA, https://www2.cito.nl/vo/share/Begrippen%20uit%20de%20TIA.pdf

Haladyna, T. M. (2004). Developing and validating multiple-choice test items (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Learning objectives for secondary education in Flanders (2023). https://www.youtube.com/watch?v=HnmYMvYyHkg

Molkenboer (2015). Toetsen volgens de toetscyclus. Deel 1. Bureau van toetsten en beoordelen, Enschede. 336 p. Enschede: Bureau van toetsten en beoordelen.

Onderwijsdoelen Vlaanderen (2024). https://onderwijsdoelen.be/

Referencing Report Flanders (2023). https://vlaamsekwalificatiestructuur.be/links-en-publicaties/koppelingsrapport-vks-eqf/files/Koppelingsrapport_BENL_2023-UPDATE.pdf

Sabbe, E., & Lesage, E. (2012). Meerkeuzetoetsen. Antwerp: Garant-Uitgevers n.v.

Sanders, Eggen, Veldhuijzen, Kleintjes, Goldebeld et al. (1994). Psychometrie in de praktijk, https://ris.utwente.nl/ws/portalfiles/portal/5593006/Psychometrie-in-de-praktijk.pdf

Teelen. (2015). Toetskwaliteit in de praktijk. Hoe maak ik goede toetsen met gesloten en open vragen? Wilp: Teelen B.V.

The matrix of secondary education in Flanders (2023). https://www.youtube.com/watch?v=8036nc2mEK4

Van Berkel, H., Bax, A., Joosten-ten Brinke, D. (eds) (2014). Toetsen in het hoger onderwijs. Docentenreeks. Bohn Stafleu van Loghum, Houten.

Verslag van de evaluatie van de examencommissie secundair onderwijs. (2022). *November-December*. https://www.vlaanderen.be/publicaties/verslag-van-de-evaluatie-van-de-examencommissie-secundair-onderwijs-november-december-2022

Vlaamse onderwijsinspectie (2024). https://www.onderwijsinspectie.be/

# Hungary

# HUNGARY: BIOGRAPHIES

## Kristóf Velkey

Kristóf Velkey started his career as a mathematics teacher and later pursued doctoral studies at the Eötvös Loránd University in Budapest and partly at the Faculty of Education, University of Warsaw in Poland with a Polish-Hungarian bilateral research scholarship. He was a member of  expert team responsible for the maths tests of the Hungarian national assessment of basic competences, and was involved in the digitalisation of the assessment. Currently, he is the head of the department responsible for analyzing national and international large scale assessment data.

## Fási Andrea

Fási Andrea graduated as an English-Geography teacher in Eger, Hungary in 1997. She taught students from age six to nineteen in public education from 1997 to 2018. She has also participated in adult education in a private language school. She has been an examiner of ECL Language Exam for years. She has always been interested in trying new methods in teaching and use digital tools during her lessons. Between 2018 and 2022 she was leading a European Union Project, which dealt with digitalism in education, finding new ways and methods of teaching making learning more enjoyable for students. Now she is at the Hungarian Education Authority where they work on digital assessment of Hungarian students in different areas like Mathematics, Reading Comprehension, Science, foreign languages, History and IT.

# FROM PAPER TO ONLINE ASSESSMENT

## Abstract

In Hungary, digital assessments have been in place since 2022. We would like to share our experiences on the reasons behind the need for and the process of digital transformation. We faced many challenges during the planning and implementation. In Hungary, we are uniquely involved in six assessment areas (Mathematics, Reading Comprehension, Science, Foreign Languages (English, German), History, IT) from this school year onwards, two of which are completely new.  Digital assessments will be carried out in eight grades, which also brings significant challenges in terms of the organization and implementation of the assessments.

## From paper to online assessment

In Hungary the paper-pencil based national student performance assessment (the National Assessment of Basic Competences- NABC) operated by the Educational Authority since 2001, was digitalised in the 2021/2022 school year, following a development process of several years. In the present school year, 2023/2024, the Computer Based Assessment of Student Competences was organised for the third time. In this chapter, we first briefly summarise the background and process of the change of medium then we present the digital assessments and finally we look at the planned developments of the future. In particular, we will look at the procedural and content changes compared to paper-pencil based assessment and the challenges faced in the transition to digital assessment.

## How we started

The development of the National Assessment of Student Competences system was introduced in 2001 based on the experience of international assessments (TIMSS, PISA) and nationally representative sample tests. The target population of the assessment changed several times in the first years and from 2004 onwards pupils in grades 6, 8 and 10 took part in the mathematics and reading comprehension assessments. At first a school sample was used which was extended from 2008 to a comprehensive assessment of the target grades except for pupils with certain disabilities. At the same time, the so-called Student Assessment Identificator was introduced, which ensured anonymous treatment of students but allowed the monitoring of their progress. The results are reported on a common proficiency scale by assessment area, which allowed the longitudinal and cross-grade comparison of the results and the monitoring of the individual progress of pupils. The statistical methodology of the assessment is based on item response theory, using a three-parameter skill scale model to make the results comparable. The results for each grade and each year were linked using a CORE test applied to a representative sample of each grade. From the very beginning an important element of the assessment system is the background questionnaire, which is completed anonymously by the participating pupils together with their parents at their own decision. This together with the school questionnaire allows the socio-cultural background of pupils and the correlation between measured performance to be analysed at national and school level.

With the technological shift that began in the second half of the 20th century and had become almost universal by the 2010s, the digitisation of NABC seemed by then to be inevitable. There were several arguments for the digitalisation of NABC:

Around the turn of the millennium, when digital tools were less common in everyday life than they are today most research focused on the correspondence between paper-pencil based and digital assessments. However, as time went on, there were increasingly strong arguments for exploiting the potential of the digital testing environment as much as possible as everyday life and work activities are nowadays largely digital and learning-teaching processes and testing should follow these changes. The international large-scale assessments, that Hungary participates in (PISA, TIMSS, PIRLS, ICILS), moved to digital platform by 2020, and their example and experiences also provided an initiative for the change.

Another motivation came from the development of the science measurement tests in 2010s, as the need to measure the new area in a digital environment was expressed and by the end of the decade the decision was made to fully digitalise the assessment system.

The 21st century requires new (digital) skills and there is a prominent view that testing should focus on these rather than copying paper-pencil based measures. Several assessment systems have attempted to resolve this situation by seeking to make improvements that maximise the potential of the digital environment while the capability to be measured is assumed to remain unchanged in its essential elements and thus no change of scale is required. In addition, attempts have been made to develop new assessment domains fully adapted to the digital environment. In further developing the National Assessment of Basic Competences we have built on the examples of PISA, TIMSS and PIRLS.

## Objectives

When digitising the NABC the main objective was to preserve, adapt and supplement the essential elements of paper-pencil based assessment, i.e. the assessment objectives and tools of the individual areas of assessment as far as possible in order to create a modern assessment that preserves the strengths of the previous one, improves it and can be continuously developed, and produces results that are comparable with its predecessor. An important aspect was to transfer the concept of knowledge, which had been established in paper-pencil based assessments, to digital assessment: the assessment tests the competences of the students in the individual assessment area rather than a test of curricular knowledge acquisition, focusing on the application of acquired knowledge and the solution of realistic problem situations. The tests are composed of tasks that test  how students can apply what they have learned to real texts, situations and problems to be solved.

Digital testing offers many opportunities for improvement beyond the extension of assessment content. At the beginning of the development process, the introduction of digital assessment was planned to be accompanied by the introduction of adaptive assessment, which would allow for a more personalised assessment based on the participants' abilities and previous answers thus providing them with more motivating and challenging tasks and resulting in more reliable data. An individual test for each pupil compiled from a central task bank also increases the objectivity of the assessment by minimising the possibility of information exchange between pupils. Computer-based testing widens the range of data that can be collected: in addition to the answers other so-called contextual data can be collected and analysed (e.g. how much time the pupil spent on a given task) which can contribute to the further development of the assessment and provide a useful data resource for secondary analyses and research. By making the tools used by the pupils, such as the calculator, the objectivity of the test is increased. Reduction or complete elimination of the weight of open-ended tasks requiring human coding, which allows full machine (automatic) coding, works in the same direction. This also means considerable organisational and economic savings but at the same time it is a requirement for the adaptivity of the assessment.

Another additional benefit of digitalization seemed to be that the organisation and implementation of the assessment would be more economical as paper-pencil based assessments entailed significant printing, storage, packaging and transport costs. At the same time, the NABC had already established protocols which needed to be adapted to the changed circumstances. From the start it was clear that the key to the change of medium was to adopt or develop assessment software that could handle the various processes of this highly complex service from task development and test design to the administration of the assessment, data collection and feedback of results to an appropriate standard. The Educational Authority decided to develop its own assessment software.

## Launch of the Computer Based Assessment of Student Competences

A precondition for the transition to online assessment was the development of online measurement software called 'Tehetségkapu' (TalentGate). The developing process of the software took place between 2016 and 2022 in several steps. After years of preparations, the first full-scale digital assessment (involving pupils in grades 6, 8 and 10) was carried out in spring 2022. In parallel with the transition to digital testing the number of measurement areas has also expanded. In addition to mathematics and reading comprehension, which were also measured on paper, science and the first foreign language (English/German) assessment areas were introduced. While the science test made its debut in 2022 at the end of a long development process, the language test was introduced into the NABC system and converted to a digital platform from an existing test based on a simpler methodology with the initial aim of making the test adaptive, an objective that has not been achieved so far due to a lack of resources.

The most significant change compared to the paper-pencil based assessment was that while this assessment was conducted nationally, on the same day and at the same time for each grade the infrastructural limitations of the number of computers available in schools made this method of assessment impossible. The extension of the assessment period had a major impact on many aspects of the assessment. This will be discussed in more detail when the assessment is presented.

Before the assessment started the main question was how the system would cope with the daily load of tens of thousands of people simultaneously. There were problems with this during the trial tests and the test days before the assessment resulted in the system hanging and freezing, but these problems were quickly solved by optimising the IT infrastructure behind the assessment. In order to reduce the impact on the testing load, the testing period was split with the 10th grade students first followed by the 8th and finally the 6th grade students at the end of April 2022. Each school was allowed to conduct the assessment according to its own schedule, but a daily limit was set for the number of participants. There were some delays and stoppages at the start of the assessment period and some pupils had to retake the assessment but apart from the first few days the test was successful.

In the case of the paper-pencil based assessments the time between writing the test and starting the analysis of the data took several months (delivery, preparation of the test booklets for coding (scanning), coding, data cleaning). One of the aims of the transition to a digital platform was to shorten this period which was not achieved for the first assessment. Only raw data was received from the platform (the student's specific answers) so it was necessary to develop a new procedure for coding and scoring the student's answers.

## Performing the assessments on the digital platform

Pupils take part in the testing of two areas of assessment on one assessment day. Each assessment area consists of a 2x45-minute test with a total of 60-70 tasks per pupil. These include 1) core tasks linking the assessments of individual years and grades, 2) items with strong parameters (already included in the main assessment) that enable a preliminary assessment of the student's ability,

### Preliminary result feedback

The digital platform provides the possibility to access information on the performance of students in the assessment in the week following the end of the assessment period for the given grade. Automatic determination of students' preliminary results has been introduced for the 2023 assessment. The proficiency score is calculated using a simplified but approximate mathematical procedure based on the students' answers to tasks with strong statistical parameters which have been in the assessment for at least one year, and their parameters were calculated using a large number of student responses.

Once the data has been cleaned and the new tasks parameterised the final student results will be calculated. These may differ considerably from the preliminary results since the final results are determined using all the tasks. However, the common experience is that the preliminary and final results give similar student results: the average difference is 2-3% of the standard deviation of the student results (200 point standard deviation, 5-7 points average difference).

3) items tested in previous years on a small sample with final parameter setting after the assessment, 4) and some newly developed and tested items that are analysed separately from the main assessment.

Students can use the identification assigned to their unique assessment ID to log in to the assessment software and start the test version assigned to them. To avoid students sitting next to each other solving the same tasks at the same time several test versions of equivalent content and difficulty are released.

The test versions are formed by combinations of blocks which are assembled by the experts, considering the proportions of the main dimensions of the content framework (test matrix) and the appropriate distribution of difficulty, i.e. only the equivalent test versions are automatically assigned and are currently assembled in the traditional way. This method reduces - but does not eliminate - the possibility of students sitting next to each other taking the same test version. At the time of the paper-pencil assessment, the administrator distributed to the students sitting next to each other versions A and B of the same test booklet with a difference in the order of the test sections.

The testbed of the assessment software can be managed with the most basic digital skills that are considered given in the 21st century. The test starts with a short introduction to the test structure and navigation, which is read out to the students by the assessor, and then the test is launched. The public page of the assessment software contains a much more detailed guide, which allows students to try out the most basic operations (e.g. scrolling) required to use the assessment as well as the use of specific digital items and task types. The introduction to this will take place at school before the assessment period. Example tasks for each assessment area are also included.

An important change compared to paper tests is that the test material remains classified even after the test has been carried out. Previously, this applied only to the core tasks that linked the tests and to tests containing sample items which were completed on a representative sample separately from the core test. Although the publication of tests and answer keys was never intended to prepare pupils for tests in schools, transparency presumably increased the acceptability of the assessment. Though the current sample tests are an attempt to be representative of the individual assessments they are really more illustrative. A positive benefit of making the measurement more secretive is that it makes it easier to meet the increased demand for tasks due to the widening of the assessment period (which will be increased by the shift towards adaptive approaches) and to provide prior feedback to the learner. On the other hand, it has the disadvantage of significantly reducing the transparency of assessment for the parties involved in the assessment such as pupils and parents, teachers and heads of institutions, which reinforced by other factors, may reduce the motivation to assess. It is therefore important to develop a mechanism to improve the visibility of digital assessment.

## Changes in the content of assessments

In the following, we describe the changes, current features and problems to be solved in the areas of reading comprehension and mathematics in the Computer Based Assessment of Student Competences. These two areas were chosen because they look back over several decades, and the comparability of the two media was therefore considered. Some of the issues discussed (in particular the digital task types and the test interface) are more or less common to the other assessment areas.

### *The digital test platform*

During the paper-pencil assessment pupils worked in a test booklet free to do as they wished. They were given one class period, or 45 minutes, to complete a test section. The same is possible with the digital assessment, and a panel with all the tasks in the test section allows students to keep track of any unsolved tasks and go to the task at any time with a click. A difference may be however, that while moving between test sections is indeed impossible because of the digital layout, in the test booklet the learner could practically go back to previous tasks. In the 2022 assessment navigation was limited: the learner could only navigate between items belonging to a stimulus but could not move back to it once he had moved on. This was particularly the case in mathematics as while in reading comprehension there were 10-12 items per stimulus in mathematics there were usually no more than 2 items per stimulus.

An important difference between the two media is the test layout: while in the booklet the tasks followed each other linearly and the learner could only access the information on the given pair of pages at a time the vertical layout of the test surface means that the information (stimulus) forming the body of the task is always on the left and the specific task on the right and in these sections one can move independently. For example, in mathematics the learner can switch between 2-3 tasks for the same diagram with the diagram remaining on the screen. This layout is even more important in reading comprehension where there are many more questions per text and whereas in the test booklet you had to turn the page between the text and the questions in the digital test the text is on the screen all the time.

Like browsers you can create tabs in the task list and switch between them by clicking on the tabs each with a title. This makes it possible to arrange different texts and diagrams in a more transparent way or to split a long text into several parts. The test platform can also display high quality, colour graphics which can be enlarged by clicking on them.

Overall, the design of the digital test platform allows learners to easily navigate between the tasks in a given test section, track their progress (time remaining, number of tasks to be solved and solved), easily access the information needed to solve a given task (texts, diagrams) and not to spend more time on the tasks than the time allotted for the test section.

### *Digital task types*

In the paper-pencil tests, the ratio of closed-ended tasks to open-ended tasks was around 60-40%. The closed-ended tasks were collected (digitised) first by data recorders and then from the mid-2010s by scanning while the open-ended ones were centrally coded using correction keys. Originally, these were processed and analysed after data recording while later they were encoded in an electronic data collection software. Although the coding of open-ended items especially those requiring longer answers has been a significant organisational and economic overload year after year they have played an important role in assessment since higher level thinking operations such as choosing an independent solution method in mathematics or evaluating reading in reading comprehension cannot be measured with closed-ended items. Since this type of task is typically the most difficult it seemed to be a difficult way of testing pupils at the top of the ability scale as the study referred to above has shown. On the other hand, as will be seen later, it is also possible to create complex and difficult tasks from closed-ended task types in the digital test platform.

Since the 2022/23 academic year, the Computer Based Assessment of Student Competences only contains items that can be coded automatically. This doesn't mean that only closed-ended items are included in the tests as - with restrictions - open-ended items are also used: items that require the typing of only one number.

If we look at the closed-ended task types in terms of complexity the least complex is the simple choice and the drop-down menu choice, while the most complex is the category choice which involves several independent items but is evaluated as a single item. The complexity of the latter can be further increased by specifying more than 2 categories (in which case the number of sub-items is of course reduced otherwise the task becomes too difficult) or by requiring multiple choices. These include items requiring multiple choices and drag and drop items in terms of complexity. By choosing the right type of item we can therefore manipulate the complexity - and therefore the difficulty - of the task. This does not mean of course that a simple choice task cannot be difficult for example in the area of text comprehension with complex statements (answer options) and strong distractors. It does mean that tasks with longer text options increase the amount of text to be read.

It is also possible to use two or more simple task types (e.g. drop-down menu, text box) of the same or different types in one item which are evaluated together. The use of complex task types either initially complex or created by combining several simple task types was also possible in paper-pencil based assessment and has been implemented to a lesser extent than at present. While the role of complex items in the paper-pencil assessment was filled by open-ended tasks the much simpler modification of the marking of the correct answer makes these task types more adaptable to the digital environment. For example, a multiple-choice category choice in the paper assessment would probably have caused strong anxiety in the students as it was more difficult to clearly correct pen-marked answers than to click from one button to another. The ease of correction and the possibility of trying things out therefore favoured the use of more complex items in digital tests.

The digital NABC includes some of the more familiar digital task types, such as drag and drop and drop-down menus (but does not include hot text or hot spot tasks) but there are also task types used in the paper assessment (e.g. tasks requiring drawing, longer explanations) that are not currently available on the platform.  Compared to paper-pencil based tests digital tests have a wider range of task types which result in more varied tests, but the number of task types is not overwhelming, they are well separated in their operation and easy to manage.

At the same time, the ease of use of the test platform and tasks raises the question of whether this might tempt learners to do some careless work and “click through” the test. Based on the current data it seems that it is exactly the simple multiple-choice tasks which are (also) the most familiar from paper tests where the number of guesses has increased. The issue needs further investigation but it is possible that students may prefer to reward the attention of interactive response forms that make better use of the possibilities offered by the digital test environment while they may not find a paper test adapted to digital testing interesting enough.

### *Digital assessment in mathematics and reading comprehension*

As mentioned earlier it did not create a new definition of assessment when changing the medium but extended it to the digital medium by adding or revising the dimensions of assessment i.e. text types and mathematical content areas as well as the individual thinking operations. The assumption behind this was that in these specific measurements (previously in the NABC test booklet and currently in the Computer Based Assessment of Basic Competences computer interface) the students are required to perform essentially/mostly the same operations in the given assessment areas. This has been made difficult by the fact that the development of the digital content framework and the digital test environment have been partly parallel and that some of the content development priorities expressed by the experts and included in the digital content framework have not yet been implemented. As a consequence, our current digital assessments are far from exploiting the full potential of assessment whether in terms of adapting paper-pencil based tasks or in the possibilities of the digital environment. Of course, the digital task types and the digital design of the test environment detailed above are important innovations in both areas but there are still significant deficiencies.

The link between paper-pencil based and digital assessment was made by transferring the core test tasks to a digital medium. These tasks were integrated into the assessment tasks and the item parameterisation of the stable behavioural tasks was carried out using the statistical parameters of the paper-pencil based assessment tasks to parameterise the other tasks.

In mathematics the paper-pencil based tests contained items requiring measurement and drawing the digital adaptation of which is not yet technically feasible on the platform. The elimination of open-ended tasks has also affected mathematics. In the paper-pencil based assessment the student recorded his calculations during the solution of the task so that it was possible to follow his train of thought for example, to accept the result of a student who followed the correct train of thought even if he made a mistake in the calculation. Similarly, it was also possible to eliminate student responses which in the course of a wrong reasoning process inadvertently led

to the same result with the correct solution. However, in the digital interface we found, as might be expected, that the student tried to record his calculation and reasoning afterwards. Although pupils are given notepaper to solve problems there is a risk that they are less likely to use it than the space left out for this purpose in the test booklet and they also perform multi-step operations by memory making it easier for them to make mistakes - and no points are awarded for the problem if the wrong result is given.

In the area of reading comprehension we can report seemingly fewer changes and it may even seem that the assignment arrangement can make it easier for students. However, the question arises, which we have already touched on in the case of task types, whether easier handling does not affect more superficial task solving. Moreover, we do not know how much it bothers students that they do not see the amount of text to be read simultaneously.  As far as innovation is concerned, although the tab layout provides an opportunity to somewhat imitate web reading we also planned to develop a website simulation for this purpose but this has not yet been realized. In the absence of this function there is no significant difference in terms of content between the paper-pencil based and digital assessment which also contributes to the fact that the paper-pencil tests already contained texts from the Internet (e.g. articles, blog posts, calls for tenders). However, the use of new, digital text types (e.g. e-mail, chat, search engine results) is a novelty. These are typically constructed texts.

We have observed a decrease in text comprehension results, and there is an ongoing discussion, whether paper-pencil based and computer-based text comprehension measure the same ability, whether students read a text displayed on a screen in the same way as a printed text. This question also arose in relation to the Hungarian results in the case of international assessments: in the 2009 and 2012 PISA assessments, in addition to the paper test, Hungary also participated in the digital reading comprehension test, and the Hungarian students performed worse on the digital reading assessment both times.

The disposal of open-ended tasks also significantly affects the area of reading comprehension, since the assessment of a complex thinking process such as reflection on the text can be solved in a more limited way with close-ended tasks, and the independent evaluation of the text is not feasible and - compared to the field of mathematics - has already been radically reduced the scope of the application of open-ended tasks, since tasks requiring the entry of numbers occur only in particular cases in reading comprehension tests. Among the thinking operations only the retrieval of information in the text can be examined with this type of task whereas previously it was used for all thinking operations. It is perhaps not a negligible fact that with the disposal of the open-ended task type, students have to read more to solve the test; although they do not have to create their own answers they do have to read several, sometimes quite complex answer options to solve the task, and they also have to perform a less creative action by choosing the correct answer. Open-ended tasks, by requiring the student to formulate the answer independently, require a different student attitude which encourages the student to think more about the answer and this works against the superficial test solution.

Comparing all of this, it is clear that for the time being many digital opportunities remain unexploited in the assessments and minor or major shifts in emphasis occur in both areas due to the possibilities and limitations of the digital test arrangement and task types as well as the narrowing of the open-ended task type and the predominance of different forms of multiple-choice tasks. Before moving on to adaptive assessment it is necessary to study these mechanisms more intensively, as well as to redesign and prioritize content and technical developments in order to create a test that is more interesting for students and better exploits the possibilities of digital assessment.

## Summary

Following the successful completion of the 2022 digital assessment the Ministry responsible for education decided to expand the assessment: During the 2022/2023 academic year, as a first step, the assessment was extended to grades 6-11. The experimental mathematics and reading comprehension assessment for grades 4-5 was carried out at the same time. From the academic year 2023/2024 the assessment in the areas of reading comprehension and mathematics take place in grades 4–11, while science and foreign language in grades 6–11. Preparations have been made to measure the areas of history and digital culture in grades 5-11 and a full population experimental measurement is carried out this year. In all areas and in all grades - with the exception of absences and exemptions - a comprehensive assessment is carried out. This shift in emphasis had a significant impact on the digitization process: the focus shifted from the adaptive development of the system to the content development of the tests of new areas and grades and to handling the increased workload.

Computer based testing – especially through the expansion of grades and assessment areas – undoubtedly imposes a greater organizational and infrastructural burden on schools than paper-pencil based testing carried out in one day, it also allows greater flexibility in the organisation of the assessment. It also enables the absent students, or in case of technical problems to take the assessment in an other day. Another difficulty from an organisational perspective is that, although all schools must meet the minimum technical requirements of the assessment it cannot be completely excluded that there are large differences either in the equipment park or in the Internet infrastructure e.g. screen size, headset quality and especially internet speed.

Institutional and especially student motivation play an important role in terms of the validity of the assessment. The six assessment areas to be conducted in 2024 and the 3 assessment days may be burdensome for many of them in the second semester of the academic year, especially given that in some grades (almost entirely in the 8th, and affecting fewer students in the 6th and 4th grades) the high-stakes high school admissions  and the 11th-grade students may already be involved in the Matura exams that are part of the higher education entrance exam starting in May. Handling these difficulties is of mayor importance in terms of the future of the assessment.

Currently, in addition to the fine-tuning of the system another development wave is being prepared in order to make the assessment more attractive and useful for users both at the institutional and student level, despite the increased testing load. The further development of the assessment in an adaptive way seems to be the most suitable solution for this with the help of which it is possible to increase the reliability of the results at the student level and to reduce the assessment burden on the students. It is also necessary to continuously update the results feedback in order to provide reliable, comprehensible and easily interpretable data to the stakeholders. In addition, by expanding the task bank it is also possible to increase the number of published example tasks.

# Ireland

# IRELAND: BIOGRAPHIES

## Emer Delaney

Emer Delaney works as a Research Fellow at the Educational Research Centre (ERC), where she oversees the development of tests for schools in Ireland. Since 2022, her responsibilities have included aspects of the ERC Drumcondra Online Testing System (DOTS), an e-assessment platform. Previously, Emer's work has focussed particularly on literacy. From 2016-2019, she led the development of a suite of new standardised tests of reading for primary schools, normed both online and on paper. She was Ireland's National Research Coordinator for the Progress in International Reading Literacy Study (PIRLS) 2021, having also worked on the 2016 cycle. Currently, she is a member of the PIRLS 2026 Reading Development Group and contributes to an initiative of the FLIP+ international e-assessment community to develop a shared online library of test items. Her wider research interests include equity in education and the intersections of gender studies with educational research.

## Adrian O'Flaherty

Adrian O'Flaherty works as a Research Associate at the Educational Research Centre (ERC). His work focusses on aspects of the computer-based testing system ERC DOTS (Drumcondra Online Testing System), mainly around test development and new system functionalities. His primary research interests are development and delivery of online assessments, and he was part of a team that oversaw the development of the ERC DOTS online testing platform, the current version of which was made available to Irish school in 2021 and was nominated for numerous eAssessment Awards. Adrian has previously worked on several international projects (PISA and TIMSS), and also on an evaluation study of a national school's support programme (DEIS – Delivering Equality of Opportunity in Schools) aimed at addressing educational disadvantage in Irish schools. He is a member of the FLIP+ international e-assessment community which is currently developing a shared online library of test items.

## Rachel Perkins

Rachel Perkins is a Research Fellow at the Educational Research Centre. She works primarily on international large scale studies at post-primary level and has managed the national implementation of a number of such studies, including the Teaching and Learning International Survey (TALIS), Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS). Currently, she oversees the national implementation of PISA and is Ireland's representative on the PISA Governing Board. Her research interests include student well-being, academic resilience, educational disadvantage, and the relationship between students' self-beliefs and their academic achievement.

## Aidan Clerkin

Aidan Clerkin is a Research Fellow at the Educational Research Centre (ERC). He oversees studies at both primary and post-primary levels, encompassing large-scale assessments and other research and evaluation work, including a recent evaluation of Ireland's Digital Learning Framework. He has experience of both national (NAMER) and international (TIMSS and PIRLS) large-scale assessments, delivered digitally and on paper. From 2018-2022, Aidan led the development of ERC DOTS, an online platform for delivery and reporting of standardised tests to schools. With the World Bank, he has contributed to research on a range of topics for countries including Morocco, Egypt, Saudi Arabia, Nigeria, and Tanzania. His research interests include social-emotional development, student engagement and wellbeing, and their relationships with academic achievement; intervention and programme evaluations; and longitudinal research methods.

## Rachel Cunningham

Rachel Cunningham is a Research Associate at the Educational Research Centre (ERC). She works primarily on standardised, screening and diagnostic tests of mathematics. From 2016-2019, she led the development of standardised mathematics tests for primary schools in Ireland, which were normed both online and on paper. She also had a central role in the development of content for an online mathematics assessment targeted at students near the end of Grade 8. As part of this work, she was involved in the development process for the ERC's Drumcondra Online Testing System (DOTS). She has also been involved with FLIP+ in developing a shared online library of test items. Rachel has worked on international assessments (TIMSS and PISA), and more recently, on the National Assessments 2021 (a large-scale national assessment of English reading and mathematics). Her research interests include all aspects of mathematics education, particularly the diagnostic applications of assessment.

# COMPUTER-BASED TESTING IN IRELAND, 2005-2024: CHALLENGES, LESSONS LEARNED, AND FUTURE POSSIBILITIES

## Abstract

This chapter traces the use of computer-based testing in schools in Ireland in two contexts: international large-scale assessments (ILSAs) (since 2005) and national standardised testing (since 2016). Learning gleaned from Ireland's participation in computer-based ILSAs informed the development of a bespoke online platform for administering and scoring standardised tests (the ERC Drumcondra Online Testing System [ERC DOTS]). Conversely, knowledge about schools' use of ERC DOTS has informed Irish approaches in subsequent ILSA cycles.

Across ILSAs and standardised testing, recurrent challenges of computer-based testing are identified, including wide variation in: (i) education technology infrastructure in schools, and (ii) students' prior experiences using this for schoolwork. Although common to primary and post-primary settings, infrastructural challenges were most apparent at primary level while differences in students' experience were most pronounced below Grade 3. Adaptations and solutions trialled are discussed. Looking ahead, further possibilities of computer-based testing, including enhanced accessibility, additional item types, and adaptive testing, are considered.

## Introduction

Computer-based platforms have been used in schools in Ireland to deliver international large-scale assessments (ILSAs) (since 2005) and nationally-developed standardised tests (since 2016). The Educational Research Centre (ERC), based in Drumcondra in Dublin, has responsibility for the administration of several ILSAs on behalf of Ireland's Department of Education. The ERC also has a statutory function to provide standardised tests and related materials to schools in Ireland, and in recent years this function has included the development and management of a bespoke platform to support the online delivery of standardised tests (the ERC Drumcondra Online Testing System, or ERC DOTS). Experiences of computer-based testing in Irish schools in the different contexts of ILSAs and ERC DOTS have informed and influenced one another.

This chapter begins by describing Ireland's participation in computer-based components/cycles of three ILSAs: the Programme for International Student Assessment (PISA), which assesses reading, mathematics, and science among 15-year-olds; the Progress in International Reading Literacy study (PIRLS), which assesses reading at Grade 4; and the Trends in International Mathematics and Science Study (TIMSS), which assesses mathematics and science at Grade 4 and Grade 8.

Next, the ERC's experiences of developing, using, and refining the ERC DOTS platform are examined. The rationale for creating the platform is considered, along with constraints and unknowns that presented challenges. Specification requirements in relation to functionality and interface are discussed in relation to the particular context of students and schools in Ireland. Additionally, a test development study in which the same reading and mathematics content was normed online and on paper is explored in detail, as it provides initial insights into how mode differences may operate in Irish primary schools.

Finally, drawing on experiences with both ILSAs and ERC DOTS, we identify persistent challenges associated with computer-based testing in Ireland, important lessons learned to date, and possible avenues of future development.

## Ireland's experiences of computer-based testing in ILSAs

### PISA

The ERC, on behalf of the Department of Education in Ireland, has administered the national implementation of PISA since its first cycle in 2000. Ireland was among a small group of countries to be involved in the early administration of computer-based elements of the assessments. As part of the field trial in 2005, Ireland, along with 12 other countries, took part in the optional Computer-Based Assessment of Science (CBAS). The assessment, which was well-received by schools and students, was preloaded on laptops provided by the ERC and up to 20 students in each of 30 schools took part in the study as part of the field trial in Ireland. However, using externally-provided laptops was both costly and time-consuming to set up and Ireland did not participate in this assessment during the main study administration of PISA 2006 (Cosgrove & McMahon, 2005).

In PISA 2009, Ireland again took part in the optional computer-based assessment, which assessed digital reading by presenting reading literacy tasks in simulated web-based environments. In total, 19 countries, including Ireland, participated in the digital reading assessment during the PISA main study, which was carried out in addition to the print-based assessment in sampled schools. Of the 35 students in each school who were selected to take part in the 2-hour print-based assessment, 15 were randomly selected to also participate in the 40-minute digital reading assessment, which took place after the print-based assessment, usually on the same day.

The digital reading assessment was delivered via a CD-ROM, meaning that schools' own devices could be used for testing. This cycle also saw a change in Ireland's test administration procedures for the PISA assessment as a whole. While external test administrators had been brought into schools to carry out testing for the first three cycles of the study (PISA 2000 to 2006), Ireland used the school associate model in 2009, meaning that a member of the school's staff administered the test to students. This change in procedures was introduced to address falling student response rates but it also facilitated the use of schools' devices for the digital reading assessment, as each device needed to be checked before testing to ensure it met the delivery specifications. In many cases, settings on schools' devices also needed to be changed to allow the devices to boot directly from the CD-ROM, which was necessary for the test to load. In practice, not every school had the capacity to carry out the digital reading assessment using their own devices and in approximately one-third of schools, laptops were provided by the ERC.

As was the case for the PISA 2006 CBAS, the digital reading assessment was well-received by schools and there was some evidence of increased engagement with it compared to the print reading assessment (Cosgrove & Moran, 2011). However, as well as the evident wide variation in education technology infrastructure in schools, feedback from teachers who administered the tests using schools' devices indicated that the work involved in checking and changing device settings was time-consuming and unmanageable. While only a small number of technical difficulties were experienced during the main study, where these did occur, they were an added burden for school staff. For these reasons, when Ireland participated in the computer-based assessments of mathematics, digital reading, and creative problem solving in PISA 2012 (which were again carried out in addition to the print assessments in sampled schools among a subset of students), laptops were provided by the ERC to all participating schools. These laptops were configured by the ERC so that they met the requirements, and the test software was preloaded onto USBs. Ireland also reverted to using external test administrators, all of whom received training in delivering both the print and digital assessments. Technical support was provided, as needed, by the ERC.

The experience of administering the additional computer-based assessments in PISA 2006, 2009, and 2012 meant that Ireland was well-placed to understand some of the challenges associated with computer-based testing when PISA made the transition to a fully digital assessment in most participating countries in the 2015 cycle. Nevertheless, as up to 42 students per school were to receive the digital assessments, the increase in the number of devices required presented a logistical challenge. Results of a survey carried out as part of the PISA 2015 field trial in Ireland indicated that about 40% of schools would not be able to complete the assessment using their own devices, while all remaining schools would require some external laptops to supplement their school devices. Thus, external laptops were again provided by the ERC for all participating schools. These laptops were transported to each school by a technical support person, who assisted the external test administrator in setting up the devices. This additional support was required to reduce the time associated with setting up 42 laptops. Technical support personnel also assisted with any technical issues that arose during testing and uploaded students' data when the session was completed.

This approach of providing external laptops, technical support personnel, and test administrators to participating schools was also used in subsequent cycles of the study (i.e., PISA 2018 and 2022). There are a number of benefits to this model, namely that the burden of checking and setting up devices is removed from schools as no access to school devices is required and there are fewer technical issues to deal with during the testing session. Furthermore, using external laptops provided by the ERC meant that it was possible to administer the test and questionnaire directly from each laptop's hard drive, rather than a USB drive, which improved the speed at which students accessed the materials. However, providing external laptops, which are hired in each cycle, is costly and rearranging test dates if requested by schools can be logistically difficult.

The move towards online testing in PISA 2025 provides another challenge, where school's access to a reliable broadband connection must be considered. Furthermore, while schools' education technology infrastructure has likely improved in recent years, initial communication with schools suggests that there are still some schools that would have difficulty providing a sufficient number of devices for PISA testing. Thus, it seems that any transition to using schools' own devices in PISA testing in Ireland is likely to be a gradual one, with some level of external support required at least in the short-term.

## PIRLS

Ireland has participated in three cycles of PIRLS: 2011, 2016, and 2021. PIRLS was fully paper-based until the 2016 cycle, when an add-on assessment of digital literacy (called ePIRLS) was administered by 14 of the 50 participating countries, including Ireland.

The ePIRLS "projects" required students to navigate through hyperlinked informational texts in a simulated web environment and to answer questions about what they read (Mullis et al., 2015). As this content was separate to that of paper-based PIRLS, the idea was that the same students should complete both tests (Martin et al., 2015), but on different mornings, with ePIRLS second. The ePIRLS software was designed for USB delivery, but could alternately be delivered via computer hard drive.

The field trial in 2015 marked the first time that large-scale computer-based testing was piloted in primary schools in Ireland. While previous experiences in PISA had flagged the variation in education technology infrastructure at post-primary level, it became clear that this was even more pronounced at primary level. Well ahead of testing, scoping visits to schools were conducted to run a system check on any available devices and to map out the room(s) designated for testing. It was common for school devices to fail the system check, mainly due to older operating systems and/or insufficient memory. As relatively few schools had dedicated computer rooms, classrooms were typically used for testing, with multiple extension leads often required to ensure that all devices could be plugged in. Scoping visits also identified the fact that a free antivirus program widely used by schools destroyed the test software (Eivers, 2019).

Following these visits, tailored plans were developed for schools. Approaches ranged from "school equipment only", through "mix and match" (whereby school devices were supplemented with laptops and/or extension leads supplied by the ERC), to "external equipment only" (whereby all laptops were provided by the ERC). Although schools were most frequently assigned to "mix and match", this often involved just a few school devices alongside a majority of external ones. Teachers re-ran the system check closer to the test day and it was not uncommon for previously-passing devices to fail. Therefore, the number of external laptops allocated to "mix and match" schools was increased, where possible, to provide contingency. During testing, some further issues with school devices became apparent. For example, in one school with a well-appointed computer room in which most devices had passed the system check, a hardware driver due for update caused simultaneous shutdowns mid-test.

Based on the experience of the field trial, it was decided to supply laptops to all schools for the main data collection in 2016. The model was similar to that used in PISA 2015 – although, for ePIRLS, laptops were purchased rather than rented, the cost being roughly equivalent. Given the expense and the logistical challenges of setup in primary classrooms, Ireland chose the option of administering ePIRLS to a random subsample of the PIRLS students – up to 22 students in each sampled school. Overall, the main study went smoothly and all data were successfully uploaded. Post-testing, the laptops were offered for sale at a low price to participating schools. This required some additional coordination by the ERC but was well-received by teachers.

An exploration of achievement in international studies is beyond the scope of this chapter. However, it is of interest that the ePIRLS data were placed on the PIRLS scale, enabling direct comparison between the paper-based and digital reading achievement of the same students. In Ireland, students did well on both tests compared to peers internationally, and – perhaps surprisingly, given the variability of education technology infrastructure in primary schools – their average achievement on paper-based and digital reading was virtually identical (Eivers et al., 2017).

In PIRLS 2021, countries could administer PIRLS either entirely on computer (with ePIRLS tasks included in the rotation and other "traditional" PIRLS texts transposed to a digital format) or entirely on paper (without any ePIRLS content) (Martin et al., 2019). Ireland initially opted for digital testing, proposing to provide laptops again to participating schools. Unfortunately, the field trial was interrupted by the COVID-19 pandemic and resultant nationwide school closures. For the main data collection, Ireland reverted to paper-based testing as, due to infection control measures, it was not feasible for technical support personnel and laptops to move between multiple schools.

PIRLS 2026 will be administered on computer for all countries (https://www.iea.nl/studies/iea/pirls/2026). As Ireland administered the ePIRLS hypertexts in 2016 but not 2021, the 2026 data will offer a first opportunity to examine national trends over time in primary-level digital literacy.

## TIMSS

The evolution of another large-scale international assessment, TIMSS, in Ireland provides an example of the interplay between domestic and international considerations in developing and rolling out computer-based assessments. Ireland participated in the first TIMSS in 1995 at both primary and post-primary levels. After a hiatus, Ireland re-joined TIMSS in 2011 (primary only) and has participated in both the primary and post-primary components in every cycle since then (2015, 2019, 2023).

Up to and including the 2015 cycle, TIMSS was entirely a paper-based assessment. For the 2019 cycle, a digital version (known at the time as eTIMSS) was developed. eTIMSS was, to a large extent, designed to be administered as a parallel version of the paper-based assessment with many items presented in substantively the same format, although additional functionality was added relating to response options (e.g., drag-and-drop) and greater use of automated scoring. The clearest difference between the paper and digital versions was that eTIMSS included an additional interactive component designed to assess students' problem-solving skills, known as Problem-Solving and Inquiry Tasks (PSIs), which had no equivalent in the paper-based assessment. Countries participating in TIMSS 2019 could choose at the national level either to administer TIMSS on paper or to move to the digital assessment. Countries that elected to administer eTIMSS also administered the paper-based assessment to a smaller sample of students in a bridging study, which was designed to facilitate the estimation of any mode effects (i.e., differences in students' performance related to the mode of the assessment – either paper or digital). In the end, half of the 64 countries in TIMSS 2019 transitioned to eTIMSS (Perkins & Clerkin, 2020).

In Ireland, consideration was given to transitioning to eTIMSS at this time, but ultimately the decision was made to remain with a paper-based administration in 2019 with the intention of moving instead towards digital administration for the 2023 cycle. There were two main factors behind this decision.

First, from a policy perspective, Ireland had rejoined TIMSS at post-primary level (Grade 8) in 2015 following a 20-year gap. Moving to a digital version of TIMSS in 2019, with the attendant risk of mode effects, would have introduced an element of uncertainty to the estimation of trends between 2015 and 2019. This was considered particularly undesirable in the context of the renewed focus at Grade 8 and given the value of TIMSS as a means of monitoring numeracy outcomes towards the end of the Department of Education and Skills' (DES, 2011b) flagship Literacy and Numeracy Strategy, 2011-2020.

Second, from a pragmatic perspective, the possibility of a digital TIMSS in 2019 was complicated by the development of a major new suite of standardised tests for assessing reading and mathematics by the ERC (www.tests.erc.ie). As described in the second part of this chapter, these were standardised in spring 2018 and released for schools' use in spring 2019, with online versions provided via the ERC DOTS platform.  The development of both ERC DOTS and the new tests, covering multiple grade levels (including overlapping grades with TIMSS), addressed a noted need to revitalise the standardised testing options available to schools (DES, 2016) and was the culmination of several years of work involving subject experts and stakeholders from across the education system, including pilot and standardisation studies involving thousands of students, as well as significant financial and professional investment in establishing the new systems. Upon the release of the online testing platform and the new tests, concerted efforts were made from 2018-2020 to raise schools' awareness of these new resources and to encourage them to move towards using the fully updated and re-normed tests as a replacement for older, more out-of-date versions.

In this context, it was felt that introducing a separate – and at that stage, unknown – international platform for administering a test of mathematics in primary schools, at the same time as rolling out ERC's own platform for administering mathematics tests in primary schools, would have presented too great a risk for confusion among schools who would use both. At the time of making the decision to administer TIMSS 2019 on paper or digitally, there was no way of knowing how the eTIMSS platform would function in practice or the likelihood of any system problems (e.g., software crashes). As such, the risk that any problems with the eTIMSS platform could have led to reputational damage to the ERC's own online standardised tests – for example, as a result of confusion between the respective platforms/tests, or due to the development of a generalised feeling among schools that online testing is difficult or unreliable – was deemed unacceptable. For both these reasons, TIMSS 2019 proceeded in Ireland with a paper-based administration.

More recently, TIMSS 2023 has seen Ireland make the move to digital administration (no longer known as eTIMSS, as digital is now the primary modality). The main data collection in 2023 was successfully carried out following a similar model to the one described above for PISA – that is, with laptops rented by the ERC and transported to schools, and with hired technical support personnel and ERC staff deployed to set laptops up in schools and assist test administrators (generally, classroom teachers) with testing sessions. With almost 11,500 students taking part in the digital assessment across more than 300 schools at primary and post-primary levels, the procurement of these hardware and personnel resources, and the logistical arrangements required to coordinate deliveries and schedules to schools around the country, represented a significant task and made much greater demands on ERC resources and time than a corresponding paper-based study would have.

Alongside this digital main study administration, Ireland chose to implement a bridging study (required for countries moving to digital in 2019, but optional in 2023) in order to gather information on the extent of mode effects, as well as on students' test-taking behaviour across the two modalities and their views of paper-based versus digital assessment. Insights from this bridging study will be crucial to interpreting the performance of Irish students in TIMSS 2023. Initial findings from both the main study and the bridging study in Ireland will be available from www.erc.ie/TIMSS in December 2024, with secondary analyses and further findings to be published from 2025 onwards.

## Development of a bespoke online testing platform (ERC DOTS) in Ireland

### Overview, rationale and timeline

Initial work on standardised assessments to be administered online to students in Ireland began in 2011. However, ERC staff were conscious of the possibilities of computer-based testing for several years prior to this, as this approach was already being adopted gradually by PISA, as described above. Another factor that prompted this new approach to assessment was the proposed introduction of mandatory standardised testing in literacy and numeracy at Grade 8 as part of a broader Literacy and Numeracy Strategy (DES, 2011b). Although this particular proposal was never enacted by the Department of Education, work on developing an online platform had already been initiated by the ERC.

In 2017 a platform to facilitate online administration of standardised tests in reasoning, English reading, and mathematics was made available to post-primary schools. The platform was developed by an external software development company to the specifications of the ERC. The online tests were developed in-house at the ERC, including the creation of the digital version of the test items, tests, and all reporting and administration instruments. The ERC was also responsible for standardising the online tests with representative samples of students prior to release.

In 2019, following several years of development work, new versions of the ERC's primary-level standardised tests in English reading and mathematics were made available on the same online platform for grades 3 through 6. A second iteration of the online platform followed in 2020.

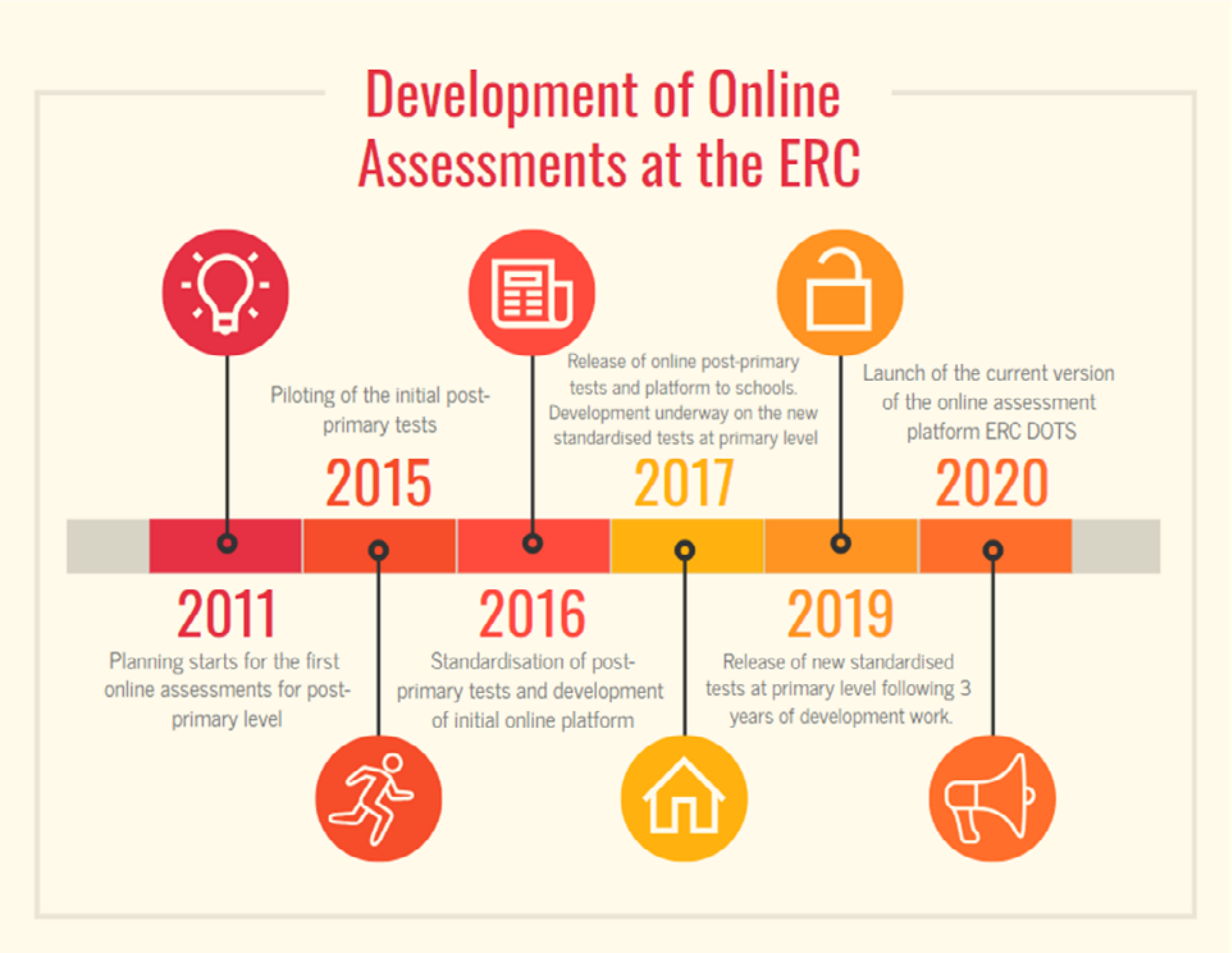

Development of Online Assessments at the ERC
2011
Planning starts for the first online assessments for post-primary level
2015
Piloting of the initial post-primary tests
2016
Standardisation of post-primary tests and development of initial online platform
2017
Release of online post-primary tests and platform to schools. Development underway on the new standardised tests at primary level
2019
Release of new standardised tests at primary level following 3 years of development work.
2020
Launch of the current version of the online assessment platform ERC DOTS

Figure 1: Timeline of the development of online assessments by the ERC

Anecdotally, there was growing demand from some schools for computer-based assessments. Online assessments were also increasingly becoming part of the large international studies, which was an additional factor behind the decision to create a computer-based option for schools for national tests.

As the technology to facilitate online assessments became more readily accessible, it was important to start exploring their obvious advantages. Among these is the reduced marking workload for teachers, with a much quicker turnaround time for reporting of results. Additionally, online assessments provide increased test security. Practising the tests in advance is not available as an option as access to the platform and to test credits for an active session is required. Replication of the online test for practice on paper would be an involved process of capturing, compiling, and printing screenshots of the content, within time restrictions (active test sessions are only available between 8am and 8pm on any given day). Online tests also allow for easier updating of content over time and more flexible piloting of new items.

On a broader level, there is increasing emphasis on the use of online resources in the classroom so assessments analogous to these pedagogical practices should be available. Digital testing platforms allow for an improved user experience with possibilities for a more engaging interface, interactive question types, adaptive testing, and specialised technology to assist test-takers where needed.

## Platform development: Specifications and design considerations

The ERC has a statutory mandate to provide and support standardised testing for schools in Ireland. In developing ERC DOTS to offer digital versions of standardised tests, the aim was to provide schools and test-takers with an easy-to-use system that is reliable, secure, engaging, and meaningful in terms of the reports that it provides. The following description of requirements and system features is based on the current ERC DOTS platform.

In developing ERC DOTS, the ERC sought to create a fully integrated online assessment system in both the English and Irish languages which would permit test development (by the ERC) and test delivery (in schools) to operate efficiently in tandem. More broadly, the vision was to create a single consolidated environment for schools to engage with all ERC tests (paper and online), thereby allowing the ERC to bring its assessment offerings to schools as one cohesive set.

ERC DOTS is viewed as a dynamic entity to help the ERC respond to changes in education policies and curricula, digital technologies, and assessment and measurement technologies. It is a flexible and responsive assessment system that helps to facilitate the ERC's strategic priorities of internal capacity building (for example, in relation to the design of new, interactive item types) and research and development (for example, enabling the use of anonymised test data to study response patterns over time).

The test development functionalities of ERC DOTS include item banking and flexible assembly of test forms. At the time of writing, there are 58 discrete online tests and approximately 3,700 discrete online test items available on the platform. A variety of section and item templates are used to suit the requirements of different tests. All such content must be manually entered on the system by ERC staff. While the current tests that are available via ERC DOTS comprise multiple-choice items, partly as a result of the parallel development of paper-based forms of the same tests (described below), additional formats (e.g., open text, drag and drop) are available within ERC DOTS for future test development.

The system is available to all schools in Ireland. Teachers can register as users, purchase online test credits, prepare for test administration by uploading test taker or class details, and carry out online testing with their students (with the ability to monitor progress in real time on a monitor portal). Immediately after testing, they can access reports of student scores, expressed as standard scores and percentile ranks on the basis of pre-established norms, using an automated scoring functionality. All administration documents (administration manual, explanatory reports, technical manuals) are available for download from the teachers' side of the system.

In addition to the ability for schools to produce reports at class and individual student level, ERC DOTS also facilitates reporting on test and item characteristics (such as classical item statistics, and test form or subscale aggregate percent correct scores) to assist ERC staff with future research and test development work.

As well as its digital assessments and related features, the current functionality of ERC DOTS includes a comprehensive payment portal to allow schools to order paper tests for delivery; a backend element to streamline ordering processes for sales staff at the ERC, facilitating efficient order preparation and tracking; and an integrated scoring tool that allows teachers to easily score paper tests and produce reports with the same detail and format as those available for the online tests.

## *Features of the online tests and the test-taker interface*

As test-takers on ERC DOTS include young children (from Grade 3 up), it was crucial that the test delivery interface be intuitive, with a minimalist design to avoid distractions on the screen while retaining all essential functionalities. It was also important to replicate (as closely as possible) the test-taking experience on paper, as parallel paper-based standardisations were occurring contemporaneously with the development of the online tests. With this in mind, online test-takers on ERC DOTS are allowed to move back and forth through the test and can review and change their answers (within the permitted time). To facilitate this, a summary review screen appears at the end of each test section. On this review screen, students can see a summary of the number of questions that they have answered, and from here they can easily navigate back to individual items and change their response if desired. The review screen also highlights the number of skipped questions and allows direct navigation to these. This feature is particularly useful for students if the allotted time is almost up.

All test instructions are available on-screen and are complemented by a read-aloud administration script. Although the option of text-to-speech is not yet available on ERC DOTS, its incorporation is planned to enhance accessibility and reduce the impact of reading load on measurement of constructs other than reading. On-screen, pre-test instructions describe the nature of each test and the user interface features needed to progress through the test. Before commencing the timed part of any test on the system, test-takers are brought through sample items of the same types they will encounter during the test. Currently, all test items on the system are in multiple choice format with four answer options, replicating the paper versions of the tests. Only one item appears on the screen at any time, allowing the test-taker to concentrate on the task at hand.

Within the test, the screen displays the test-taker's name (as entered on the system by the teacher) along with a timer at the top that counts down to show the remaining time. The title of each test section is displayed throughout, along with the item number and the number of items in the section. These features help the test-taker to orient themselves within the test and to monitor their progress. Additional test interface features to enhance engagement include child-friendly, colourful illustrations and a user-friendly format (e.g., comprehension texts in reading tests are presented across multiple tabs to minimise scrolling).

### *Technical specifications and operational requirements*

To ensure the longevity and robustness of ERC DOTS, the system has been designed to be compatible with API (Application Programming Interface) and QTI (Question and Test Interoperability specification) standards. This allows maximal uncoupling of the platform's structure and content features to minimise risk associated with interdependencies, and to permit flexibility in future development (e.g., easier incorporation of potential future modules for scoring of text responses, provision of an offline assessment solution, or adaptive testing).

Being mindful of the variation in levels of IT infrastructure and internet access across schools (as described earlier in the context of ILSAs, and also evident from other research – see for instance Feerick et al., 2021), local device and network requirements have been kept to a minimum. An additional feature is a dummy test which allows schools to test their devices' ability to run tests prior to a test session (https://trythetest.erc.ie).[1]

Minimising the risk of server overload and maximising stable and complete response data, while taking account of changing numbers of users and fluctuations in local connectivity during testing, were prominent considerations in the development of ERC DOTS. At the same time, a key consideration, influenced by the varying quality of internet access in schools, is the ability for ERC DOTS to recover effectively from temporary local or server-level crashes or disruptions, with no loss of data and minimal impact on test-takers. For example, following a temporary drop in internet connectivity or if a student has to log out mid-test for any reason, upon logging in again the test-taker is returned to the last response recorded, without loss of data and with the correct time remaining on the test as if there had been no interruption.

Finally, it was considered essential that ERC DOTS would be accessible and intuitive for teachers to use. This requirement applied across a range of functions, including ordering tests, setting up test sessions, administering tests, accessing reports, and accessing supporting documentation. With this in mind, templates are provided within ERC DOTS for the upload of test-taker details, while report templates are downloadable in several formats (e.g. individual reports, class reports) and can then be uploaded to schools' existing content management systems (CMSs) without the need for any modifications. Since launching the platform, some work has been done towards further increasing compatibility with common CMSs (e.g., to allow for direct importation of class lists from a CMS and direct exportation of reports to a CMS).

## Case study: Development of primary school tests in two modes (2016-2019)

This section describes how the ERC's online testing platform was used during a large-scale test development project.

[1] Many of the features of the test-taker interface as described in this section can be viewed by readers at trythetest.erc.ie.

## *Background*

Since 2012, it has been mandatory for primary schools in Ireland to conduct standardised testing of English reading and mathematics at grades 2, 4, and 6, to report results to parents, and to report aggregated results to the Department of Education (DES, 2011a).[2] The ERC provides tests in these subjects for grades 1–6 and, in practice, many schools opt to test at grades 1, 3, and 5 as well. The development of new versions of the primary tests was prompted in part by recognition that the norms for previous versions, standardised in 2005 and 2006, had become outdated (DES, 2016). This was probably due to a combination of schools' increasing familiarity with the test content and a genuine improvement in reading and mathematics standards in Irish primary schools, as also observed in national and international assessments (National Assessments of Mathematics and English Reading [Shiel et al., 2014], TIMSS [Clerkin et al., 2016], and PIRLS [Eivers et al., 2017]). Recognition of this issue signalled an urgent need for the tests to be redeveloped, with a new set of norms,  so that students' performance could be described relative to an up-to-date population.

A key decision involved the mode(s) through which the redeveloped tests should be made available. Ireland's experience of administering ePIRLS in 2016 had highlighted the variability of infrastructure in primary schools. Moreover, while ePIRLS had not required internet access, the ERC's DOTS platform did, and it was anticipated that poor broadband might prove an additional barrier to digital testing. Therefore, continuing to provide a paper-based version of the new tests was essential to ensure accessibility for all schools. The question that remained was whether providing online versions of the tests as well would be of substantial benefit for those schools that had the infrastructure to avail of them.

In considering this question, one unknown was the extent to which primary school children in Ireland were familiar with using digital devices for schoolwork or similar purposes. If familiarity was low, or variable, this might impact on their performance on tests of reading and/or mathematics – i.e., there would be a substantial mode effect, which was not desirable. The findings from PIRLS 2016, whereby students in Ireland had performed equally well, on average, on paper-based and digital reading, were tentatively encouraging in this regard. However, in PIRLS 2016 questionnaires, many students in Ireland indicated that they had learned their computer skills mainly outside the classroom (Eivers, 2019). Furthermore, it was far from certain that the behaviours of Grade 4 PIRLS students would apply across all the grade levels targeted by the new tests.

Weighing these concerns against the advantages of online testing for schools, including the reduced burden for teachers and the potential for the interface to improve children's experiences, it was decided to pilot the new tests on both paper and online modes for grades 2-6. This would allow for data on mode effects to be collected at all these grade levels – a first in Ireland – which would inform the approach taken in the subsequent standardisation.

[2] In primary schools in which Irish is the medium of instruction (about 8%), standardised tests of Irish must also be conducted and reported on at these grade levels.

### *Pilot (2017)*

For convenience, pilot schools were selected on the basis that they had at least two classes per grade level and, per their responses to a survey, had at least 25 usable digital devices and high-speed broadband. The intention was that, at a given grade level, one class in a school would take the tests on paper and the other class would take them online. In practice, some schools that had indicated that they had sufficient infrastructure ended up needing additional laptops, and sometimes Wi-Fi routers, to be supplied, while a few schools could not participate in online testing at all. More than 2,800 students in 56 schools took part in paper-based testing, while more than 3,300 students in 52 schools took part in online testing.[3] At each grade level, multiple forms (versions) of each test were piloted.

Pilot item statistics showed that the majority of items in both reading and mathematics were easier on paper than on computer, although there were exceptions. At the level of test forms, mode differences in the difficulty of the reading tests ranged from 0-6%, while mode differences in mathematics ranged from 0-8%. At lower grade levels, mode differences tended to be larger and more consistently indicative of test content being more difficult on computer, particularly in mathematics. While the overall trend suggested that the online format was a little more challenging in both subjects, especially for younger students, the differences observed were considered small enough to allow the tests to proceed to standardisation in both modes. As part of the post-pilot review, items with very large mode differences were removed or adjusted, while minimising mode differences at form level was a consideration when items were rearranged to balance the forms. Finally, given the larger mode differences observed among younger students, it was noted that the suitability of both modes for grades 2 and 3 in particular would be reviewed post-standardisation and prior to release.

In preparation for the standardisation, significant changes were made to the Grade 2 mathematics test. In the pilot, the full test was read aloud to students, meaning that the same form of the test had to be administered within each class. However, informal feedback from schools indicated that this approach was problematic when administering the online test as copying was facilitated by the proximity and visibility of neighbouring screens. To a lesser degree, copying was also a concern for paper testing. In response to this feedback, the Grade 2 pilot forms were reworked as parallel forms for the standardisation. The first block of items was read aloud, but with minor variations in the numbers, images, and answer options used in each form, resulting in different correct answers. For the latter section of the test, the students worked through the items while reading independently, with different content in each form. This allowed multiple forms to be administered within the same class group. At more senior grade levels, although the mathematics test was not read aloud, teachers had the option of reading any words with which a student had difficulty, to reduce the impact of the reading load. This came with the proviso that mathematical terms could not be defined.

[3] The larger number of students in fewer schools for online testing reflected the fact that schools were offered the option to test additional classes online if they wished. Illustrating some of the advantages of online testing, this created minimal additional costs for the ERC, whereas if schools had tested additional classes on paper increased costs would have been incurred for printing, postage, data entry, and shredding.

### *Standardisation (2018)*

The standardisation brought new challenges. It was important for the tests to be standardised in both modes on a representative sample of the population – which included students in schools with and without the relevant education technology infrastructure. Therefore, unlike in the pilot, schools were sampled without regard to the availability of suitable devices or broadband. Each sampled school was asked to conduct paper-based testing at four grade levels and online testing at two grade levels.[4] Based on participating schools' reports of their own resources, a majority were provided with at least some laptops by the ERC, and about half were provided with a Wi-Fi router on the test day(s). There were proportionately more technical problems in the standardisation than had occurred in the pilot, often due to poor internet connection in schools.

Anecdotally, in schools where laptops were provided, some Grade 2 students in particular were observed struggling with practical aspects of the computer-based tests such as logging in, using a mouse to select answers, navigating between items and tabs, and zooming to enlarge content. Students at the same grade level who took the tests on their own school's devices (often tablets) appeared to have fewer such difficulties. Item statistics demonstrated that, in the representative standardisation sample, there was a substantial mode effect for students at Grade 2. As seen in the pilot, students at this grade level found the same items harder on computer than on paper; however, the differences at test form level in the standardisation were large and systematic enough in both subjects to be of concern. The overall percentage of correct responses on paper was broadly in line with what was expected and targeted. On the other hand, the lower percentage of correct responses online suggested that students who took the test on computer had had a non-comparable, and possibly demoralising, test-taking experience. This was especially the case in the reading test, where two out of three Grade 2 forms were about 8% more difficult online than on paper, with the remaining form about 5% more difficult online. On closer inspection, the relative difficulty of the online format at this level was especially noticeable when students were required to zoom to read text, and/or when content was distributed across tabs in such a way that each tab contained a semantically discrete section (for example, a specific character's perspective).

At other grade levels, form-level mode effects ranged from small to negligible. In reading, forms were slightly more difficult on computer at grades 3-5, but less difficult on computer at Grade 6. However, the patterns were somewhat different for mathematics. At Grade 3, one form was more difficult on computer, while the other was marginally less difficult. At grades 4 to 6, none of the forms were more difficult on computer, although there was some variation in the extent of the apparent advantage to taking the test on computer. It may be, then, that children's interaction with the platform was not consistent across the two subjects. This does not seem altogether surprising as the test format differed considerably by subject. For example, the reading tests included sections featuring both an item pane and a text pane, with the latter split across multiple tabs, whereas the mathematics tests featured just one pane.

[4] This was considered a reasonable trade-off between the need to gather data from a sufficient number of students on each mode versus the cost and logistical challenges of supporting less-equipped schools to conduct online testing.

Having reviewed the magnitude of mode differences across grade levels, and giving due consideration to the benefits of online testing for schools that could avail of it (e.g., a reduced administration and marking burden for teachers), it was decided to release the tests for grades 3-6 in both modes but to release the Grade 2 tests in paper format only. A somewhat analagous decision had already been taken in relation to the response format for paper tests: while students at grades 3–6 marked their answers on machine-scorable answer sheets separate to the test booklets, students at Grade 2 marked their responses directly into their test booklets as the use of answer sheets was deemed likely to cause too much construct-irrelevant variance among this age group.

The paper-based and digital formats of the grade 3–6 tests were scaled separately, so that the norms now used by schools compare students' performance with that of peers in the standardisation sample who took the same test via the same mode.

## Conclusion: Lessons learned and future possibilities

For schools, computer-based testing has been a paradigm shift from traditional methods. Therefore, there is an existing culture that has needed to adapt in order to take advantage of the conveniences and possibilities of digital assessment and keep pace with student engagement in an online world. This change has needed, and continues to need, careful management and system-level support to ensure that all stakeholders benefit. To facilitate this change on a wider scale, a robust ICT infrastructure (including reliable internet access) is needed across all schools in Ireland, with more work required at primary than at post-primary level. Contingent factors for an improved ICT landscape are ongoing professional development, technological knowledge, access to technological assistance, and embedding of the use of digital technology among school staff (Cosgrove et al, 2022; Donohue et al, 2024; Feerick et al, 2022). Continued learning and updating of skills are also required by ERC staff working in test development, particularly as the field of digital testing continues to evolve. Additionally, data security considerations are constantly growing and it is important for the ERC to remain up-to-date with best practice and relevant legislation, such as the European Union's General Data Protection Regulation and Ireland's Data sharing and Governance Act.

To date, the main model used to administer computer-based ILSAs in Ireland has involved the ERC renting or purchasing laptops and transporting these to schools for testing. This has facilitated Ireland's participation in important global developments in large-scale assessment and the collection of useful data regarding students' proficiency in a digital environment. However, it is a model that is very costly, as well as administratively burdensome – for schools as well as the ERC. Many of the infrastructural impediments encountered during the administration of computer-based ILSAs also presented challenges during the development of ERC DOTS. In developing the platform, it was possible to specify some design features intended to adapt to the national context – for example, the ability of the programme to run on a wide range of devices and to respond flexibly to crashes due to loss of internet connection.

Nevertheless, developing online tests for schools that have insufficient or variable education technology infrastructure inevitably involves constraints. The simultaneous development of analogous paper and digital assessments (to satisfy the needs of all schools) places heavy demands on the ERC's resources. It also requires the content and format of online tests to mirror closely those of their paper-based equivalents, restricting the ability to take advantage of interactive item formats. Additionally, mode effects may be more exacerbated in the norms than in the data resulting from use of purchased tests, as the norm group is based on a sample drawn to be representative of students in all schools while the online tests are typically purchased by a self-selecting set of schools with the necessary infrastructure.

The ERC is at the forefront of computer-based assessments in Ireland and will continue to use the knowledge and experience gained to date (through involvement in ILSAs, collaboration with national and international colleagues, and with development of ERC DOTS) to support schools in the transition to this mode of assessment. Additional possibilities of the digital format for national test development are now being explored, particularly in the areas of adaptive testing, innovative item types, and enhanced accessibility. Contributing to this, the ERC is a member of the FLIP+ international e-assessment community, a not-for-profit entity comprising researchers from many countries whose goal is to share experience and build solutions to enhance assessment globally (flip-plus.org). This includes working groups focussing on a broad range of assessment topics (e.g., process data, psychometric data, inclusion and accessibility). Much of the work of FLIP+ to date has centered on the development of an international item library which will facilitate the sharing of ideas, test items, and technical knowledge among its members.

In Ireland, there are several ongoing government strategies and frameworks in place designed to support the country's 3000+ primary schools and 700+ post-primary schools in their use of digital technologies (Cosgrove et al., 2019). In particular, the Digital Strategy for Schools to 2027 (Department of Education, 2022) – which succeeded a previous Digital Strategy running to 2020 (Department of Education and Skills, 2015) – has as one of its core pillars the improvement of education technology infrastructure in schools in Ireland. The commitments in the Strategy range from providing funding to schools for the purchase of digital technology to enhancing high-speed broadband connectivity and Wi-Fi in schools. For example, €210 million was provided to schools via an ICT Infrastructure grant between 2015 and 2020, with a further €200 million investment and an additional €13 million to improve schools' broadband connectivity promised for the period to 2027 (Donohue et al., 2024). The need to improve technical support services to schools and to streamline procurement frameworks is also acknowledged, as well as the need to provide guidance and advice to schools in areas related to digital technology. However, while improvements have been made in recent years, significant challenges in terms of connectivity and the availability and suitability of digital devices in schools are still apparent, especially among primary schools (Donohue et al., 2024).

A common principle underpinning all these supports, and another key pillar of the Digital Strategy to 2027 (Department of Education, 2022) is the promotion of the embedding of digital technologies and digital pedagogy in the classroom. Researchers at the ERC have recently completed a national, longitudinal evaluation of one such initiative, the Digital Learning Framework (Cosgrove et al., 2019; Donohue et al., 2024). Nonetheless, challenges remain also with regard to teachers' use of digital technologies, especially at primary level (Feerick et al., 2022).

As the development of digital literacy skills becomes increasingly prioritised in policy in Ireland (Department of Education, 2024), there is a growing need for assessments of digital literacy, and therefore some of the challenges noted here become more pressing. Notwithstanding these challenges, the willingness of schools to take part in computer-based ILSAs and the increasing number of schools using ERC DOTS demonstrate a positive inclination towards digital assessment. As we review progress to date and look towards the future, we recognise the importance of consultation with stakeholders – particularly teachers and students – to help guide future developments in e-assessment in Ireland.

## References

Clerkin, A., Perkins, R., & Cunningham, R. (2016). *TIMSS 2015 in Ireland: Mathematics and science in primary and post-primary schools*. Educational Research Centre. https://www.erc.ie/wp-content/uploads/2016/11/TIMSS-initial-report-FINAL.pdf

Cosgrove, J., Feerick, E., Moran, E., & Perkins, R. (2022). *Digital technologies in education – Ireland in the international context: Trends and implications from PISA 2012-2018.* Educational Research Centre. https://www.erc.ie/wp-content/uploads/2022/09/DT-in-Ed-PISA-2012-to-2018-for-web.pdf

Cosgrove, J., and McMahon, S. (2005). *Report on PISA field trial of the Computer-Based Assessment of Science: Operations and outcomes in Ireland.* Educational Research Centre. https://www.erc.ie/documents/p06ftcbas_nat_report.pdf

Cosgrove, J., and Moran, G. (2011). *Taking the PISA 2009 test in Ireland: Students' response patterns on the print and digital assessments.* Educational Research Centre.

Cosgrove, J., Moran, E., Feerick, E., & Duggan, A. (2019). *Digital Learning Framework (DLF) national evaluation: Starting off – Baseline Report.* Educational Research Centre. https://www.erc.ie/wp-content/uploads/2020/01/DLF-national-evaluation-baseline-report.pdf

Department of Education and Skills. (2011a). *Circular 0056/2011. Initial steps in the implementation of the National Literacy and Numeracy Strategy.* https://assets.gov.ie/13643/037a8791893f466fa4bf439c47f6d8fa.pdf

Department of Education and Skills. (2011b). *Literacy and Numeracy for Learning and Life: The National Strategy to Improve Literacy and Numeracy among Children and Young People 2011–2020*. https://assets.gov.ie/24521/9e0e6e3887454197a1da1f9736c01557.pdf

Department of Education and Skills (2015). *Digital Strategy for Schools 2015-2020: Enhancing Teaching, Learning and Assessment.* https://www.gov.ie/pdf/?file=https://assets.gov.ie/25151/52d007db333c42f4a6ad542b5acca53a.pdf#page=null

Department of Education and Skills. (2016). *Standardised achievement tests: An analysis of the results at primary school level for 2011-12 and 2012-13.* https://www.gov.ie/en/publication/c27a52-standardised-achievement-tests-an-analysis-of-the-results-at-primary/

Department of Education (2022). *Digital Strategy for Schools to 2027*. https://www.gov.ie/pdf/?file=https://assets.gov.ie/221285/6fc98405-d345-41a3-a770-c97e1a4479d3.pdf#page=null

Department of Education (2024). *National Strategy for Literacy, Numeracy and Digital Literacy, 2024-2033*. https://www.gov.ie/pdf/?file=https://assets.gov.ie/293255/a509a8d7-a4ac-43f9-acb0-29cdc26a1327.pdf#page=null

Donohue, B., Moran, E., Clerkin, A., Millar, D., O'Flaherty, A., Piccio, G., & Dinh, T. (2024). *Digital Learning Framework (DLF) national longitudinal evaluation: Wave 2 Final Report*. Educational Research Centre. https://doi.org/10.70092/0412063.0824

Eivers, E. (2019). *Left to their own devices: Trends in ICT at primary school level*. Irish Primary Principals' Network. https://issuu.com/ippn/docs/left_to_their_own_devices_final

Eivers, E., Gilleece, L., & Delaney, E. (2017). *Reading achievement in PIRLS 2016: Initial report for Ireland*. Educational Research Centre. https://www.erc.ie/wp-content/uploads/2017/12/PIRLS-2016_inital-report-IRL.pdf

Feerick, E., Clerkin, A, & Cosgrove, J. (2022) Teachers' understanding of the concept of 'embedding' digital technology in education. *Irish Educational Studies*, 41(1), 27-39. https://doi.org/10.1080/03323315.2021.2022521

Feerick, E., Cosgrove, J., Moran, E. (2021). *Digital Learning Framework (DLF) national evaluation: One year on – Wave 1 Report*. Dublin: Educational Research Centre. https://www.erc.ie/wp-content/uploads/2021/06/DLF-W1-full-report.pdf

Martin, M.O., Mullis, I.V.S., & Foy, P. (2015). Assessment design for PIRLS, PIRLS Literacy, and ePIRLS in 2016. In I.V.S. Mullis & M.O. Martin (Eds.), *PIRLS 2016 Assessment Framework* (2nd ed.), pp. 55-69. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2016/downloads/P16_FW_Chap3.pdf

Martin, M.O., von Davier, M., Foy, P., & Mullis, I.V.S. (2019). PIRLS 2021 Assessment Design. In I.V.S. Mullis & M.O. Martin (Eds.), *PIRLS 2021 Assessment Frameworks*, pp. 57-74. Boston College, TIMSS & PIRLS International Study Center. https://pirls2021.org/frameworks/wp-content/uploads/sites/2/2019/04/P21_FW_Ch3_AssessDesign.pdf

Mullis, I.V.S., Martin, M.O., & Sainsbury, M. (2015). PIRLS 2016 Reading Framework. In I.V.S. Mullis & M.O. Martin (Eds.), *PIRLS 2016 Assessment Framework* (2nd ed.), pp. 11-29. Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2016/downloads/P16_FW_Chap1.pdf

Perkins, R, & Clerkin, A. (2020). *TIMSS 2019: Ireland's results in mathematics and science.* Educational Research Centre. https://www.erc.ie/wp-content/uploads/2021/01/03-ERC-TIMSS-2019-Report_A4_Online.pdf

Shiel, G., Kavanagh, L., & Millar, D. (2014). *The 2014 National Assessments of English Reading and Mathematics. Volume 1: Performance Report*. Educational Research Centre. https://www.erc.ie/wp-content/uploads/2016/11/NA_2014_Vol1_Final-updated.pdf

# Kosovo

# KOSOVO: BIOGRAPHIES

## Osman Buleshkaj

**Osman Buleshkaj** pursued graduate studies in Canada and Slovenia focusing on educational leadership and knowledge management. Osman has been a high school teacher, university professor and leadership trainer. He has been a consultant for teaching, leadership and education policy with various national and international organizations supporting education reforms in Kosovo in the last 15 years. He currently works as an Associate Researcher with the Kosovo Pedagogical Institute.

## Fatmir Elezi

Fatmir Elezi  completed his university studies at the Faculty of Education in the subject of mathematics. For three years he worked with the lower secondary school as a mathematics teacher. For 21 years, he has been working in the assessment division at MESTI in national student assessments. He has been engaged in the role of data manager for the international studies PISA, TIMSS and PIRLS. He currently holds the position of the Head of Division for assessment standard and monitoring at MESTI.

## Selim Mehmeti

Selim Mehmeti pursued graduate studies in University of Pristina, master's degree on educational management. He worked as a high school teacher, researcher, leadership trainer, and educational official in the Ministry of Education. He coordinated and contributed to the development of curriculum for gymnasiums, guides, manuals and by-laws for the implementation of the Strategy for Quality Assurance in pre-university education, Strategy for Teacher Development in Kosovo. His area of expertise is training of school principals for management of education and quality assurance. His research covers the themes of educational policy, advancement of leadership in education in Kosovo, teacher development, curriculum implementation, quality assurance, assessment in education. He published many professional articles for the research and analysis in fields and issues of interest to education in general.

# E-TESTING AND COMPUTER-BASED ASSESSMENT IN KOSOVO

## Abstract

As Kosovo invests in improving its digital infrastructure and widening access to digital resources, it paves the way for adopting e-testing and computer-based assessments. In recent years, e-testing and computer-based assessments have been used in Kosovo during international assessments such as PISA and TIMSS. These implementations faced challenges related to infrastructure, access, and preparation. Additionally, various schools in Kosovo have adopted online platforms for computer-based assessments, more about the circumstances created in the conditions of the COVID 19 pandemic. However, these efforts currently lack cohesive national policy guidance.

This study aims to investigate the current state of e-testing and computer-based assessments in Kosovo, focusing on infrastructure, access to technology, training provisions, and an analysis of the benefits, challenges, and potential for application. A mixed-methods approach will be employed, sampling policy makers, staff responsible for planning and organizing national and international assessments, educators, and representatives of donor organizations.

While e-testing and computer-based assessments offer promising avenues for advancing Kosovo's education system and aligning it with international standards, successful implementation depends on addressing challenges in infrastructure, access to training, and policy.

## 1. Introduction

The incorporation of technology in education has transformed traditional teaching approaches, offering more dynamic and interactive learning experiences. It is widely acknowledged that integrating digital technology in the education system is an asset for both teachers and students. The use of high-quality equipment such as computers, tablets, projectors, televisions, and so on, can facilitate the learning process and directly contribute to preparing students for more successful careers in an increasingly digital future.

In their review, Shute and Rahimi (2017) suggested that computer-based assessment (CBA) for learning (CBAfL) will contribute to improving personalized learning in a variety of contexts, and that the innovative CBAfL techniques will move beyond the laboratory works into more practical applications in many subjects. However, they claim that application of CBA will overcome boundaries between instruction, learning and assessment so the need for high-stake tests of learning will become unnecessary. While different education systems advance in CBAfL, teachers will have to be trained to provide targeted support and personalized learning for diverse students, continues and formative assessment will replace formal exams, and that students will be equipped with the knowledge and skills needed to succeed in the 21st century (Shute & Rahimi (2017).

Technology has created new learning opportunities, thereby enhancing the effectiveness of teaching and learning. The internet has granted students access to a wide range of information and resources from around the world, encouraging them to improve their research skills. Another significant advantage of using technology in schools is the availability of virtual labs and various educational games designed to help students grasp different concepts in certain subjects, thereby enhancing the learning experience by making it more interactive (Haleem A., et al, 2022). Furthermore, technology has provided students with new opportunities to collaborate and communicate, allowing them to work together on various project assignments regardless of their location. On the other hand, technology has simplified classroom management for teachers by enabling them to use digital tools such as online materials or textbooks, to assess and grade assignments, to keep records of student participation, and to monitor their progress. These tools save teachers time and make the learning process more efficient (Camilleri, M.A., Camilleri, A.C., 2017).

The first volume of PISA results from the OECD explores the use of technology for learning and its relationship with PISA scores (OECD, 2023). The findings from PISA 2022 indicate that students who spent up to one hour per day using digital devices for learning activities at school, regardless of their socio-economic background, achieved mathematics scores that were 14 points higher than their peers. This positive association was observed in more than half of the education systems included in the analysis, comprising 45 countries and economies with available data (OECD, 2023).

According to the PISA 2022 Results, students are confident in using digital technology for distance learning, but they still prefer to learn autonomously. For instance, across all OECD countries, approximately three out of four students reported feeling confident or very confident about using learning management systems, digital learning platforms, or video communication programs, but they also expressed the need for guidance and support from teachers. This highlights that merely providing training on technological tools is insufficient; students must also be prepared to take responsibility for their own learning (OECD, 2023).

The average PISA test scores of Kosovo students with access to a digital environment at home were consistently higher than those without, across all subjects. Regarding the frequency of learning activities conducted with digital resources, over 80% of teachers reported using digital resources for many learning activities. Common practices included utilizing online tools for student assessment and providing access to learning materials for students unable to attend classes physically. However, most teachers indicated that they implement these activities only once or twice a month (OECD, 2023). Therefore, it is essential for the Ministry of Education and supporting institutions to take actions and assist teachers and schools to effectively utilize digital technology. Consequently, such support would help build capacities for implementing e-testing and computer-based assessment in Kosovo.

There is a shortage of digital equipment in Kosovo schools, with computers, projectors, digital boards, printers, and internet devices being among the most essential. Even in schools with sufficient devices, many are outdated, slow, or partially non-functional due to a lack of ongoing maintenance. The use of technology presents a challenge due to inadequate training of teachers. Data from a report published by KCDE for Prishtina municipality indicates that in classes where technology is used for teaching activities, students perceive greater involvement from their peers, and teachers are eager to incorporate technology more frequently (KCDE, 2023).

The education system in Kosovo primarily relies on traditional methods of learning, with limited integration of technology (KPI, 2020). Schools still report an insufficient number of computer labs and limited internet connectivity. Despite these challenges, teachers found that online learning not only served the purpose during the emergency closure of schools but also had a positive impact on their attitudes toward integrating digital components into their traditional teaching methods (Morina et al, 2020). Therefore, to present a brief overview of e-testing and computer-based assessment practices, this report shall provide an analysis of relevant education policies, the capacity building of teachers in digital technologies, examples and experiences from classroom applications, and strategic interventions planned in the following years.

## 2. Relevant educational policies

Educational policies, as well as procedures and approaches to student assessment in Kosovo, have been analyzed and discussed, focusing on the reference elements related to electronic testing and computer-based assessment. In this process, the reference elements include relevant educational policies that are in the process of implementation, such as strategic documents, curriculum provisions for pre-university education, and instructions for student assessment in the pre-university education sector.

*Education Strategy*. The Education Strategy is the main document for the development of the education sector in Kosovo in the period 2022-2026 (MESTI, 2022a), there are five strategic objectives, defined for the five priority areas. The student assessment component is addressed in Strategic Objective 2: Raising the quality of pre-university education through the consolidation of quality assurance mechanisms and the provision of quality teaching. Within this objective, the creation of the Center for Evaluation and Standards is foreseen to ensure sufficient human capacities that will increase the reliability of the tests.

While within the framework of Strategic Objective 5 - Digitization of education, in addition to the creation and functionalization of a comprehensive digital platform for the field of education, a special component is also foreseen related to the development of digital competence in function of the successful digital transformation of education, which also includes the development of digital competence among teachers in the field of assessment, namely the use of digital strategies and technologies for improving assessment.

The Education Strategy does not envisage any special and direct measure for the organization of electronic testing and computer based assessment, at the same time it does not prohibit it but let it be understood that this can be achieved based on the progress for the Digitization of education.

*Curriculum provisions*. The curriculum system of pre-university education consists of the conceptual document defined in the Curriculum Framework for Pre-University Education (CFPE) in Kosovo (MESTI, 2016); the Core Curricula for the formal levels of education including primary education (MESTI, 2017a), lower secondary education (MESTI, 2017b), and upper secondary education (MESTI, 2017c). For each of the education levels, the specific subject curricula are provided for every curriculum field and grade level.

In relation to student assessment, the CFPE document describes the general goals, principles, and types of assessment to ensure consistency and consistency of the student assessment system. The CF encourages the balanced use of different assessment approaches for systematic monitoring and evaluation of students, moving towards competency-based assessment (MESTI, 2016).

CF defines two types of student assessment, internal assessment at the school level and external assessment by the central authority for assessment authorized by the Ministry of Education. According to the CFPE, internal evaluation is done at the school/class level by schoolteachers and according to the description of the procedures and criteria for each type of internal evaluation, regulated by by-laws. Whereas the external assessment is a standardized assessment to measure the level of achievement of learning outcomes, mastery of competencies at the end of level I, II and III of pre-university education.

Based on the definitions set out in the CFPE document for the learner assessment component, breakdowns for the details of learner assessment go further into the Core Curricula for the formal levels of education and the subject curricula for each grade.

None of the curriculum documents directly describe or define the organization of electronic testing and computer-based assessment. At the same time, curriculum documents do not prohibit this form of evaluation and indirectly allow the possibility of such approach to improve and advance the evaluation process in general (MESTI, 2016).

*The Framework for Assessment*. The Framework for Student Assessment (FSA) in pre-university education level in Kosovo presents a coherent and comprehensive description of how internal assessment, external assessment, and international assessment are organized and integrated (MEST, 2020). The document describes in detail:

- General characteristics of student assessment.
- Student evaluation procedures, internal evaluation, and external evaluation.
- Student evaluation capacities.
- Approaches to building capacities for student assessment.
- Reporting and using student assessment findings.
- Future developments in national student assessment.

In addition to the descriptions above, the assessment framework in separate chapters deals with the aspects of teacher evaluation and the aspects of school evaluation, in connection with the evaluation of students. While in a separate chapter, the evaluation of the educational system is dealt with in connection with the role of student evaluation.

The Framework for Student Assessment (FSA), among other things, provides guidelines for internal assessment and external assessment of students. In relation to the internal evaluation, the framework instructs teachers to respect the requirements of the curriculum for the evaluation of students, according to which: (i) the evaluation of students must be guided by the evaluation principles; (ii) the main focus of internal assessment should be to support students' learning to master the competencies and this is best achieved by the combination of formative assessment (for learning) and summative assessment (of learning); and (iii) internal assessment should enable all students to express new knowledge and show the level of competence mastery.

As for the external assessment, the FSA regulates the assessment of students at the national assessment level happening at the end of 5th grade, 9th grade and the Matura Exam at the end of 12th grade. The framework for national assessments instructs the central authority for assessment authorized by the Ministry of Education to develop follow-up instructions for the preparation of assessment requests/questions. Instructions include guidelines for preparation of test models, reporting forms at the end of national assessment, and information for the general public and the educational community at large (MESTI, 2020).

Like other documents, the FSA for pre-university students in Kosovo does not directly describe and guide the organization of electronic testing and computer-based assessment. However, the framework promotes the use of empirical data and supports the transition to evidence-based decision-making, so that it is understood that this data can also be provided through the organization of electronic testing and computer-based assessment.

In addition to the Framework for Student Assessment, the Ministry of Education has adopted a bylaw on assessment in pre-university education which is based on principles of transparency, impartiality and trustworthiness (MESTI, 2022b). For the purpose of this study, we analyzed the possibilities to conduct electronic assessment as regulated by this bylaw. It can be stated that electronic assessment can fulfill these principles very well, and it provides for a good opportunity to create an electronic evaluation system that fully supports the contemporary requirements for quality and successful evaluation. However, implementation of e-testing and computer-based assessment,  is challenge that requires institutional support, relevant infrastructure, as well teacher training and capacity building programs.

## 3. Trainings for teachers

According to the educational legislation in Kosovo, the policies of professional development of teachers are directed by the Ministry of Education. Ministry is responsible for regulating the system of career development of teachers, creating mechanisms for the implementation of professional development of teachers, drafting standards for ensuring quality, but also for accreditation of in-service teacher training programs.

Programs for teacher training in the field of student assessment were identified and analysed from the catalog of accredited and approved programs for the professional development of teachers and education leaders (MESTI, 2022/23). The emphasis of the analysis was placed on the argumentation that teacher training programs in the field of student assessment include the development of teachers' competencies related to assessment approaches and the practice of electronic testing and computer-based assessment.

Analysis of the catalog of accredited and approved programs for the professional development of teachers and educational leaders, fifth edition (MESTI, 2022/23), showed that out of a total of 115 thematic programs for teachers and education leaders, 4 of them are directly related to the field of student assessment:

- Reading assessment in the early grades.
- Summative assessment of students: Designing the test.
- Assessment of students based on evidence.
- Online assessment and ongoing student support.

Training programs, early grade reading assessment, and evidence-based student assessment do not address and provide elements related to assessment approaches and the practice of electronic testing and computer-based assessment.

Meanwhile, the training program: ***Summative Student Assessment: Test Design***, addresses and provides approaches and practices related to assessment that can be applied in the classroom, but also through electronic testing and computer-based assessment. Among other things, the training content focuses on the topics below:

- Developing the test and linking the questions/requests to the learning outcomes for the field and competencies for the curricular level.
- Test table, standard procedures for designing a test, scoring, and returning points to the grade.
- Designing questions according to Bloom's Taxonomy for knowledge levels, improving questions, test content.
- Types of questions, instructions for writing questions, questions with completion, questions with alternatives, questions with association, questions with graphic presentation, questions with structured answers, rules for writing them.
- Problems, the structure of their solution. Designing questions/requests and linking them to learning outcomes for domain and competency for degree.
- Analysis of the test/questions. Question analysis, question difficulty, question discrimination, alternative frequency.

E-testing and computer-based assessment issues are addressed directly and more comprehensively in the Programme, Trainer: Basics of Online Learning offered by the local organization Kosovo Centre for Distance Education (KCDE), respectively in the module: Online Assessment and Ongoing Learner Support. To illustrate this, in the following section we present three topics of the online assessment module as provided by the KCDE:

**Online assessment: An introduction**

Topics to be covered:

- Main evaluation methods.
- Analyzing and discussing examples of online assessment.
- Formative and summative assessment methods and strategies.
- Presentation of Bloom's taxonomy.
- Analyzing and discussing assessment questions.

**Assessment in LMS**

Topics to be covered:

- Understanding the main features of a quiz.
- Creating assignments in Moodle.
- Creation of multiple-choice tests.
- Creation of a series of evaluation methods in Moodle (quizzes, rubrics, etc.)
- Using different question formats (open/closed questions; multiple choice)
- Creating a quiz in Moodle.

**Use of online assessment platforms**

Topics to be covered:

- Reading materials on common online assessment platforms.
- Exploring the Socrative platform with a step-by-step video.
- Setting up a quiz.
- Using one of the assessment platforms and creating a quiz.

To otherwise empower the usage of the LMS, a comprehensive Training of Trainers (ToT) program of Instructional Designers got developed and conducted by the Kosovo Centre for Distance Education (KCDE). The 5-weeks long TOT program covered various aspects of digital education, including components, methodologies, and the use of Moodle e-learning tools and techniques. In addition, a 2nd edition of the ToT program is further intended to start later in 2024. The team should thus be empowered to offer the program beyond this implementation period on its own. To ensure quality of learning content on its LMS, a digital content framework was additionally developed and introduced. Currently, there are around 30 participants attending the second edition of ToT. This shall contribute to building capacity for future trainings in this field.

The online assessment module and the continuous support of students is a good opportunity for teachers to develop competences in the field of assessment, particularly for assessment approaches and the practice of electronic testing and computer-based assessment. However, this program, from the duration of the accreditation and the possibility of offering the involvement of teachers in this training program, does not ensure the minimum of raising the capacities of teachers at the level of the system to apply electronic testing and evaluation through the computer.

## 4. Examples of the application of e-testing

**At the school level.** There are several platforms that are present in Kosovo, and schools occasionally use them to facilitate the implementation of the lesson, such as: School.Me.education, Busulla.com, Teacher gaming platform, and Millennium 3 online e-learning. However, digital learning has more the meaning of the integration of technology in the classroom, and not the integration of technology in different subjects or grades. There are examples from some schools that develop forms of online learning in an individual way, as well as apply forms of e-testing and evaluation through computers.

A report provided by Millennium 3 school (https://mileniumi3.net/publikime/ ), a non-public schoo*l, shows that e-testing is applied to improve the assessment process through digital pla*tforms, especially Google Workspace for Education and Moodle. Trainings are organized for the use of Moodle, Google Workspace for Education, AI as well as new technologies, and e-courses have been created and made available to all school staff.

Google Workspace for Education is used as the primary platform, where all learning materials, assignments and student projects are placed. Within it, Google Forms is used in the evaluation process. Teachers create personalized quizzes and tests using the various options, which can include different types of questions such as multiple-choice, short-answer, and essay-type questions. The ease of use and approach of Google Forms has improved the assessment process, allowing teachers to create, distribute and grade tests efficiently.

In addition, the school uses Moodle as a second platform. On this platform, each teacher has created his own e-course. Within this course, they create quizzes and tests, which can be customized with different question types and settings. In addition, they also use Moodle's powerful features that support secure test administration, using the Safe Exam Browser. Implementation of e-testing has brought many benefits for both students and teachers of this school. Students appreciate the convenience and flexibility of electronic testing, allowing them to complete assessments at their own pace and receive immediate feedback on their performance. Teachers, in turn, have seen improvements in assessment efficiency, assessment accuracy and access to secure data. E-testing has also facilitated a seamless transition to distance learning during times of disruption, especially during the COVID period.

**At the Ministry of education level.** Evaluo.ORG was an all-inclusive cloud-based testing platform created in Kosovo, and it supported online creation and delivery of professional, feature-rich tests. The platform came with its own apps for using on IOS and Android devices as well as with its responsive web interface for using on desktops. ‘Evaluo’ provided support for all testing processes: from test creation, test delivery, candidate management, results reporting and analytics. This platform provided support for more than 15 different question types, which could be used for easy creation of tests, as well as question banks.

Question banks served as repositories for questions of different categories, types, as well as difficulty levels. For each created question, authors could also include reference to syllabus, learning materials or other references. In this way they could create tests to be used as diagnose tests or preparatory tests. Test authors had the option to publish tests as public, private, or only for themselves. In case of public tests, which were accessible through everyone, test authors could choose to share the news through all social media and different communication channels.

The delivery of private tests was supported by the Exam-Manager, which allowed scheduling of exams, creation or selection of candidates, invitation of candidates, as well specification of additional exam characteristics, such as shuffling of questions, bulk scheduling, manual scheduling, retakes, number of times test can be taken, surfing through tests, etc. Test taking for candidates was very easy. In case of public tests, candidates could take the test in their preferred device (mobile, notebook, tablet, or desktop). Upon finishing the test, platform automatically would generate the results, which can be analyzed by candidate, or test authors. And finally test authors could generate detailed reports and use excellent analytics.

‘Evaluo’ was developed in accordance with the best practices and fulfilled all the required standards for web and mobile applications. The technologies used for ‘Evaluo’ were: Java SE, Swift, PHP, Laravel, MySQL, ReactJS, Ajax, HTML5, CSS3. **API Application Program Interface –** modern API was built with Ajax, PHP7 and Laravel 5.5 and came with the highest level of security currently available (Passport Oauth2). **Android OS –** The android application was built in the JavaSE programming language. The application’s architecture was MVC - Model View Controller and the database for storing local data was used Shared-Preferences and SQLite. **iOS –** The iOS application was built in the Swift. For the storing of local data, the User Defaults and SQLite was used, and application’s architecture was MVC. **WEB –** the web interface was developed by using the latest technologies, such as HTML5, jQuery, CSS3, JavaScript, etc.

A study was conducted by Thaçi (2019) on application of e-testing for Matura exam at the end of pre-university education in Kosovo. To advance the Matura Test process at the national level, the Ministry of Education in Kosovo planned to conduct the Matura Electronic Test in 2017. An application named MATU application was created with electronic tests for students. The purpose of the application was to prepare the students for the electronic Matura exam so that they would be familiar with how an electronic Matura assessment works. By downloading the application, they would take the tests prepared for eight subjects. The tests had similar design and similar questions with the Matura exam. After they had completed the tests, the application enabled the students to receive the results at the same time immediately after the test was completed. The idea was to increase their interest and motivation to attend the Matura exam electronically.

The MATU application preparation test contained 24 tests, 8 for grade, as follows: 8 for grade 10, 8 for grade 11 and grade 12 for the following subjects: Mathematics, Physics, Chemistry, Biology, History, Geography, English and Mother tongue. This study (Thaçi, 2019) investigated the use of MATU application and its effect on students‘ interest to being evaluated electronically. This analysis included only the cases of students who downloaded the application and completed the test. In 2017, there were 24,152 high school graduates in Kosovo of which 7,332 or 29.59% of graduates registered to take MATU test online.

Taking into consideration that the launch of this application was made without supportive information campaign and considering the high interest of graduates to participate online, it was considered that Kosovo could move towards the full digitization of Matura test under the condition that relevant technologies are made available to schools, teachers were trained, and students were informed timely and accurately (Thaçi, 2019). It is worth noting that that this process did not continue due to technological and budgetary implications it entailed.

## 5. Conclusions and recommendations

Even though the period of the COVID 19 pandemic has influenced the rise of this need, it has not been taken seriously by the institutional mechanisms of the education system in Kosovo. Participation in trainings related to application of technology and computer-based assessment continues to be driven by program providers’ capacities and not Ministry of Education funding and support. There are several platforms that are currently utilised in Kosovo, and schools occasionally use them to facilitate the implementation of the lesson. In addition, the Ministry of Education had a project to develop online resources for application of e-testing and computer-based assessment, and a pilot Matura exam was conducted online. However, e-testing and computer-based assessment are still not developed and not integrated into teaching, learning and assessment practices in Kosovo. Digital learning in Kosovo has more the meaning of the integration of technology in the classroom, and not the integration of subjects or classes into technological tools. In cases where schools choose to use teaching and learning alternatives, they use various platforms and online teaching programs. As such, there are examples from some public schools and private schools that developed forms of online learning and assessment, which the central education institutions could learn from. Kosovo has a five-year strategic plan to digitalize the education system and work to improve the infrastructure required for application of technology in teaching and learning.

## References

Camilleri, M.A., Camilleri, A.C. Digital Learning Resources and Ubiquitous Technologies in Education. Tech Know Learn 22, 65–82 (2017). https://doi.org/10.1007/s10758-016-9287-7

Haleem et al., (2022) Understanding the role of digital technologies in education: A review, Sustainable Operations and Computers,Volume 3, Pages 275-285, ISSN 2666-4127, https://doi.org/10.1016/j.susoc.2022.05.004

KCDE (2023). Digital School Mapping in the Municipality of Pristina through Focus Groups. Prishtina, XK: funded by UNICEF office in Kosovo. Retrieved from https://kcde-ks.org/hartezimi-digjital-ishkollave-ne-komunene-prishtines-permesfokus-grupeve/

Kosova Pedagogical Institute (2020). Mësimi në distancë/e-mësimi në arsimin parauniversitar në Kosovë, në rrethanat e krijuara nga pandemia Covid-19: Përmbledhje e hulumtimit [Distance education/e-learning in pre-university education in Kosova under the circumstances created by the Covid-19 pandemic: Research summary]. Retrieved from https://ipkmasht.rks-gov.net/wp-content/uploads/2021/02/Mesimi-ne-distance-E-mesimi-ne-arsimin-parauniversitar-ne-Kosove-ne-rrethanat-e-krijuara-nga-pandemia-covid-–-19-2020.pdf

Ministry of Education, Science, Technology and Innovation (MESTI) (2016). *Kosovo Curriculum Framework*, MESTI, Prishtina, RKS. Available online at: https://masht.rks-gov.net/korniza-kurrikulare-e-arsimit-parauniversitar-te-republikes-se-kosoves/

Ministry of Education, Science and Technology (MEST) (2017a) *Core Curriculum for Education Level 1*, MEST, Prishtina, RKS. Available online at: https://masht.rks- gov.net/uploads/2017/02/kurrikula-berthame-1-finale-2.pdf

Ministry of Education, Science, Technology and Innovation (MESTI) (2017b) *Core Curriculum for Education Level 2*, MEST, Prishtina, RKS. Available online at: https://masht.rks- gov.net/uploads/2017/02/korniza-berthame-2-final_1.pdf

Ministry of Education, Science, Technology and Innovation (MESTI) (2017c) *Core Curriculum for Education Level 3*, MEST, Prishtina, RKS. Available online at: https://masht.rks- gov.net/uploads/2017/02/korniza-berthame-3-final.pdf

Ministry of Education, Science, Technology and Innovation (MESTI) (2020). The Framework for student assessment in pre-university education in Kosovo. Prishtina, XK: MESTI. Available online at: https://masht.rks-gov.net/korniza-e-vleresimit-te-nxenesve-te-arsimit-parauniversitar-te-kosoves-2/

Ministry of Education, Science, Technology and Innovation (MESTI) (2022a). The Education Strategy 2022-2026. Prishtina, XK: MESTI. Retrieved online from: https://masht.rks-gov.net/wp-content/uploads/2022/11/03-Strategja-e-Arsimit-2022-2026-Eng-Web.pdf

Ministry of Education, Science, Technology and Innovation (MESTI)  (2022b). The administrative instruction nr. 06/2022 on student assessment in pre-university education of the Republic of Kosovo. Retreived online from: https://masht.rks-gov.net/wp-content/uploads/2022/09/UA-PER-VLERESIMIN-E-NXENESVE_Final-ISNIU-e-derguar-tek-SP-per-nenshkrim-00000002h.pdf

Morina M., Uka, A., Raza, K. (2021). A Case Study on Kosovan Teachers' Transition to Distance Education during COVID-19 Pandemic, International Journal on Innovations in Online Education. https://doi.10.1615/IntJInnovOnlineEdu.2021038933

OECD (2023), *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/53f23881-en

Shute, V. J. & Rahimi. S. (2017). Review of computer-based assessment for learning in elementary and secondary education. *Journal of computer assisted learning. Vol 33*(1), p. 1–19. https://doi.org/10.1111/jcal.12172

Thaçi, L. (2019). The impact of the matu application on the students' interest in Kosovo. A paper presented at 14th International Balkan Congress for Education and Science in Ohrid, North Macedonia. Retreived online from: https://www.researchgate.net/publication/343451952_THE_IMPACT_OF_THE_MATU_APPLICATION_ON_THE_STUDENTS'_INTEREST_IN_KOSOVO

# Luxembourg

# LUXEMBOURG: BIOGRAPHIES

## Dr. Philipp Sonnleitner

Philipp Sonnleitner is a psychologist with expertise in assessing students' competencies, abilities, and attitudes. He currently serves as the Method Lead for Luxembourg's national school monitoring tests known as Épreuves standardisées. His personal mission in research and teaching is to develop valid, high-quality assessment methods, using the latest statistical and computer-based technology to ensure their quality, and to give guidance concerning their correct and fair application in educational contexts. In the course of his career he has been involved in developing tests and assessment procedures for the school monitoring programs in Austria and Luxembourg, as well as the OECD's PIAAC-, and PISA assessments. He also gathered extensive experience developing instruments to capture competencies ranging from college aptitude to complex problem-solving behavior.

## Steve Bernard

Steve Bernard is a researcher in the field of educational testing and develops mathematics assessments for the Luxembourg Centre for Educational Testing (LUCET). He has a master's degree from the Faculty of Psychology at the University of Luxembourg. In addition to his studies in the field of psychological intervention, he has been a student research assistant in the domain of media psychology for over 6 years. He is currently particularly interested in the research topics of automatic item generation and the gamification of math tests.

## Dr. Sonja Ugen

Sonja Ugen is head of the Luxembourg Centre for Educational Testing (LUCET) which is mainly responsible for the implementation, enhancement and assurance of the Luxembourgish school monitoring programme Épreuves Standardisées (www.epstan.lu). LUCET is further involved in large scale assessment projects such as cognitive and language testing, university admissions testing and student course feedbacks, many of which are computer-based using LUCET's in- house online assessment system OASYS. Coming from a background in psychology, Sonja Ugen is furthermore implicated in the development of diagnostic tools allowing to screen for or diagnose (developmental) disorders in a linguistically heterogenous school population.

# EMERGING TRENDS IN E-ASSESSMENT: INSIGHTS FROM OASYS AND THE IMPACT OF GENERATIVE ITEM MODELS

## Abstract

This chapter provides a comprehensive exploration of e-assessment, delving into its multifaceted advantages and challenges. It begins by tracing the developmental path and critical insights gathered from the utilization of the OASYS (Online-Assessment SYStem) assessment platform, a cornerstone of educational practices in Luxembourg. Subsequently, the focus shifts towards a critical aspect prevalent in all e-assessments: the creation of test content. Regardless of the intended application, whether it is used for adaptive testing, formative assessments, or summative evaluations, the necessity of robust and psychometrically sound test content remains paramount. Within this context, the chapter illustrates the innovative approaches adopted by the Luxembourg Centre for Educational Testing (LUCET) in addressing this challenge. Specifically, it highlights the implementation of template-based, generative item models in a large-scale mathematics assessment conducted nationwide. Furthermore, the chapter explores the growing interest in generative artificial intelligence (AI) and its potential implications in this context. Through a nuanced examination of these themes, this chapter offers valuable insights into the current trends and future directions of e- assessment.

## Introduction

When the international PISA study switched the primary mode of assessment to computer-based administration in 2015, it was clear that this was the gold standard for large-scale assessment and e-assessment in education (OECD, 2016). Remarkably, this was already predicted (and much sooner expected) in the 1980s by the US-based Educational Testing Service (ETS, Bunderson et al., 1988). Back then, the future of e-assessment[1] looked bright, and the authors anticipated that computer-based assessment will soon not merely administer test items that look identical to paper-pencil based tests (substitution) but will expand them with digital components (transposition), and finally also diagnose and adaptively test students' abilities (transformation). In hindsight, those projections – although being sharp observations of the huge potential of e-assessment – were overly optimistic given the current state of e-testing that is mostly still dealing with transposition and transformation (Fischbach, Greiff, Cardoso-Leite, & König, 2021).

Using the so-called technology hype cycle model from tech consultancy Gartner (Gartner, 2018),

[1] Note that in the remainder of the chapter, we use the term e-assessment as umbrella term for assessment administered on electronic devices (i.e. computers, tablets, or smartphones).

the slow development comes as no surprise. Each innovation therefore passes five stages to reach full productivity, including the initial innovation, the peak of inflated expectations, the trough of disillusionment, the slope of enlightenment, and the plateau of productivity. According to this model, Bunderson et al.'s 1988 paper could definitely be seen as the peak of inflated expectations concerning e-assessment and the decades that followed let us realize the resource-hungry development, the often cumbersome handling of user interfaces, potential security threats, a clear lack of concepts for technology-enhanced item formats, and insufficient item pools to maximize potential ways of testing of and for learning.

However, the integration of Artificial Intelligence (AI) into e-assessment systems holds promise for overcoming these challenges. AI can enhance the adaptability of assessments, allowing for personalized learning experiences tailored to individual student needs (Miao, & Holmes, 2023). Additionally, AI algorithms can analyse vast amounts of data to identify patterns and trends, facilitating more efficient item development and improving the overall quality of assessments (Miao, & Holmes, 2023). Nonetheless, it is essential to approach AI integration thoughtfully to address concerns regarding quality, data privacy, bias, and ethical implications (Holmes et al., 2022; Le Borgne et al., 2024).

In this particular chapter, we are discussing challenges of e-assessment platforms in relation to their form, and content – a distinction that is made across various fields, but especially for web applications. We are describing two cases, based on how the Luxembourg Centre for Educational Testing (LUCET) responds to these challenges: (a) optimization of form through usability testing and UX design, and (b) generation of content through model-based item generation.

## Description of case

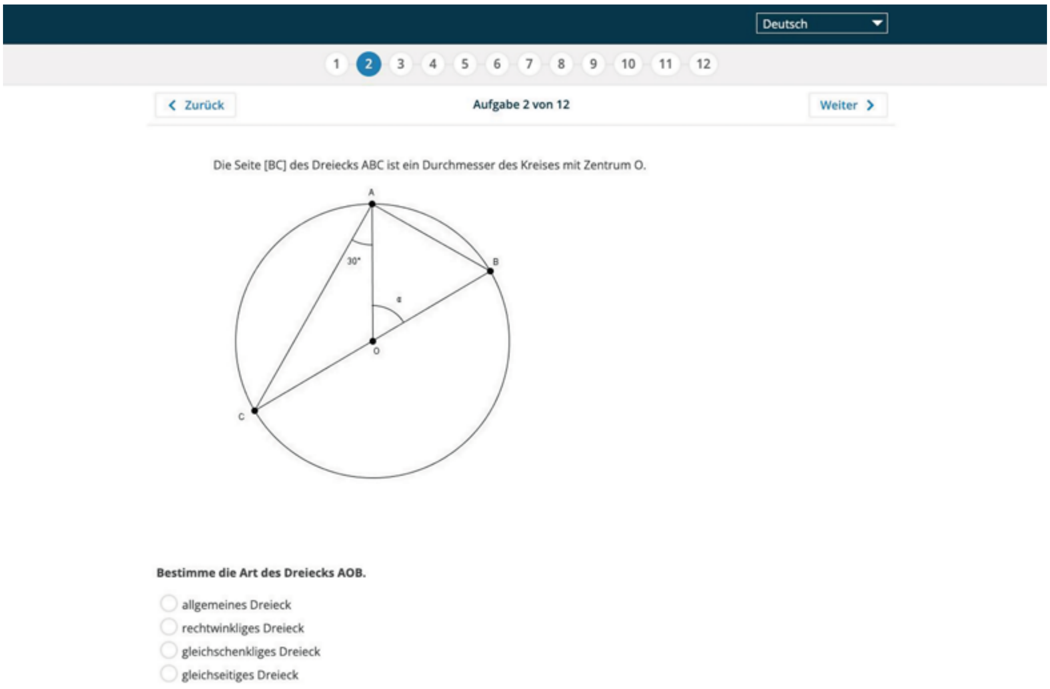
### Luxembourg's response to e-assessment demands: the online assessment system (OASYS)

Due to the many advantages of e-assessment, Luxembourg's school monitoring programme Épreuves standardisées (ÉpStan, cf. epstan.lu) decided to administer secondary school tests through web-based platforms from the very beginning. Relatively quickly, however, it became clear that off-the-shelf solutions didn't meet LUCET's expectations regarding test design, test security, technological reliability, and ease of use. As a consequence, in 2010, it was decided to develop an in-house testing and exam platform called OASYS (Online-Assessment System) that allows for easy building and delivering of tests (Fischbach, Greiff, Cardoso-Leite, & Koenig, 2021)

Currently, OASYS effectively addresses various substitution scenarios (i.e. administering traditional test and questionnaire formats), and extends its capabilities beyond mere transposition tasks (i.e. making use of the digital environment for assessment formats). For instance, it incorporates innovative interactive elements like digital concept mapping. OASYS is reliable, as data is immediately transmitted and stored, and connection interruptions are instantly detected and displayed in a related surveillance mode. By offering easy navigation throughout the entire test and fast loading of items, it is pleasant to use for the test-taker (see Figure 1 for an example mathematics item). OASYS also provides access to behavioral data, allowing for the tracking of actions such as switching between displayed items and languages, as well as recording answers even if they are later changed (indicating either an initial error of the student or insecure response behavior). Leveraging the expertise of LUCET in assessment alongside the former human-computer interaction research group at the University of Luxembourg, the development of OASYS prioritized user-centric development to optimize user experience (for both, test developers as well as test takers), ensure superior data quality, and enhance learning processes from the data (Fischbach et al., 2021).

As previously highlighted (Sonnleitner et al., 2017; Sonnleitner, 2019), the extensive effort invested in crafting the GUI (graphical user interface) or UX (user experience) is well-founded. In today's educational landscape, students hold specific expectations regarding technology. They anticipate flawless functionality, influenced by their exposure to high-quality commercial computer programs or applications. They also seek intuitive interfaces, drawing from their experiences with video games and modern mobile devices. Complexity in navigation is irritating; interfaces are preferred that are easy to understand without the need for extensive instructions. The GUI should be visually appealing, aligning with contemporary design standards (as experienced in everyday use of tablets, smartphones, etc.), to enhance the perception of test quality. In addition, students prefer to learn through active exploration and interaction, rather than through lengthy written instructions, so it's best to provide example items during instructions. Failure to meet these expectations may jeopardize the acceptance of e-assessment among students.

To maintain these high standards of GUI and UX for both, test item creators and test-takers, the system was mainly developed internally at LUCET. This allowed for direct and transparent communication, and immediate feedback loops that helped identifying and addressing issues and bugs.



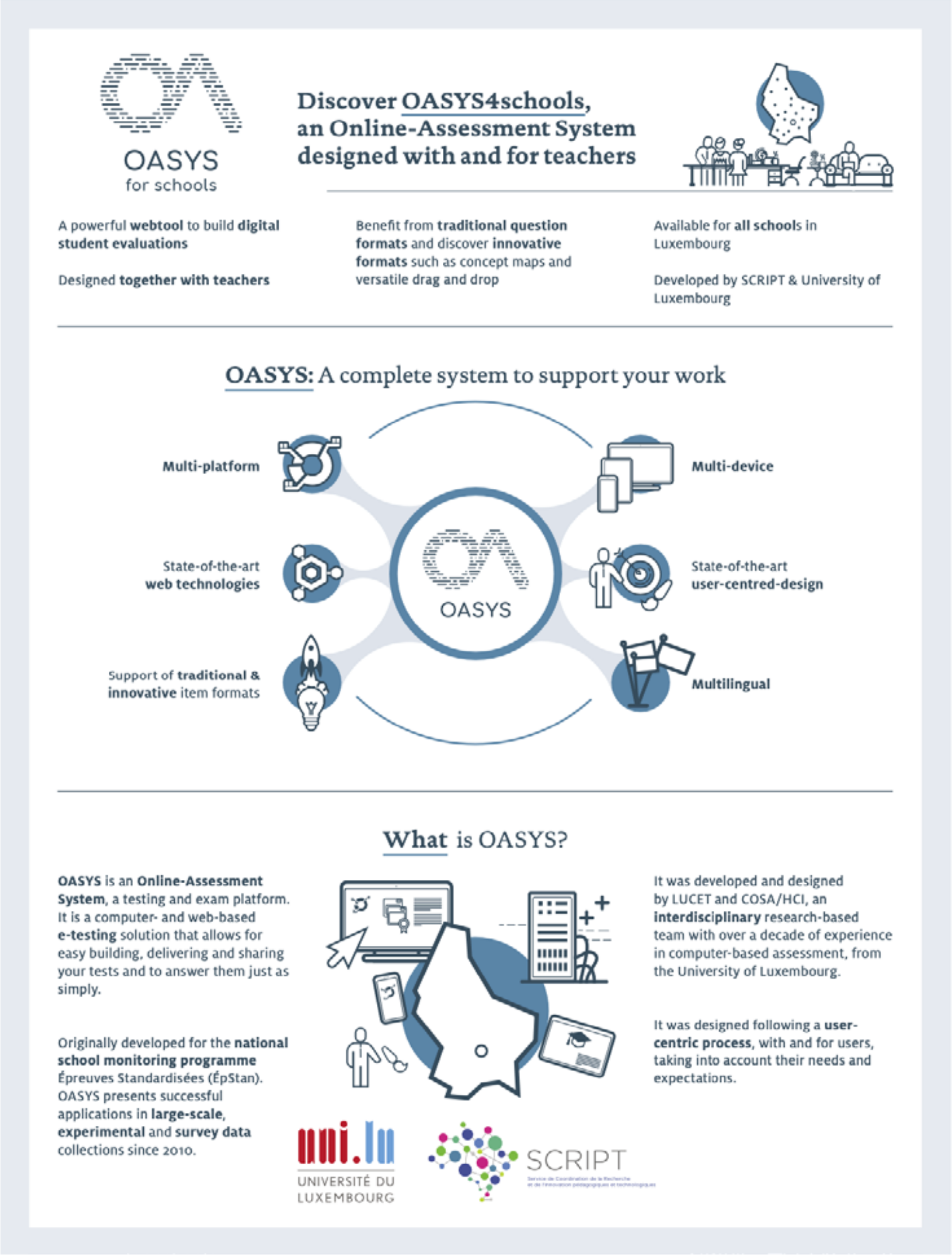
Die Seite [BC] des Dreiecks ABC ist ein Durchmesser des Kreises mit Zentrum O.

**Bestimme die Art des Dreiecks AOB.**

○ allgemeines Dreieck
○ rechtwinkliges Dreieck
○ gleichschenkliges Dreieck
○ gleichseitiges Dreieck

**Figure 1**: Screenshot of a mathematics test item delivered in OASYS. Administration language can be switched in the upper right corner. The navigation pane in the upper center indicates the position within the test and whether an item was already responded to or not.

In 2018, due to its versatility and high usability, OASYS was officially made the standard e-assessment and e-exams platform for Luxembourg's educational landscape. This joint cooperation "OASYS4schools" (Fig. 2; oasys4schools.lu) between the SCRIPT (Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques, i.e. the ministry of education's division for pedagogical and technological innovation and quality assurance) and LUCET ensures a continuous user-centered development of the platform that considers the demands of the field and latest innovations of research at the same time.



**Discover OASYS4schools, an Online-Assessment System designed with and for teachers**

A powerful **webtool** to build **digital student evaluations**

Designed **together with teachers**

Benefit from **traditional question formats** and discover **innovative formats** such as concept maps and versatile drag and drop

Available for **all schools** in Luxembourg

Developed by SCRIPT & University of Luxembourg

**OASYS:** A complete system to support your work

**What is OASYS?**

**OASYS** is an **Online-Assessment System**, a testing and exam platform. It is a computer- and web-based **e-testing** solution that allows for easy building, delivering and sharing your tests and to answer them just as simply.

Originally developed for the **national school monitoring programme** Épreuves Standardisées (ÉpStan). OASYS presents successful applications in **large-scale**, **experimental** and **survey data** collections since 2010.

It was developed and designed by LUCET and COSA/HCI, an **interdisciplinary** research-based team with over a decade of experience in computer-based assessment, from the University of Luxembourg.

It was designed following a **user-centric process**, with and for users, taking into account their needs and expectations.

**Figure 2**: Overview on OASYS4schools demonstrating the broad and user-centric development approach of the e-assessment platform (taken from Fischbach et al., 2021)

## The way OASYS is used and its further potential

In OASYS, a comprehensive array of methodologies exists for crafting assessments tailored to the diverse needs of children and teenagers. Examples for implemented tests range from German, and French reading comprehension tests, to mathematics, concept map building, and questionnaires covering a broad range of topics. The most frequently used way of test delivery involves linear tests (sequence of items is fixed), which offer structured evaluations that may incorporate branching to accommodate varying proficiency levels response trajectories. Branching in this case is not (yet) done by calculating sum scores, but by predefined pathways based on the chosen answer option(s).

An alternative way of testing is realized by OASYS' so-called fluid tests: Pools of items are implemented and defined based on specific characteristics (e.g. measuring the same sub-competency or same difficulty level). The fluid testing method then randomly picks a predefined number of items of each pool to finally compose unique linear tests. This approach mostly is applied when items are known to share the same psychometric characteristics and item exposure should be kept low.

Expanding beyond linear and fluid testing, OASYS also facilitates the creation of multiple linear tests, allowing test creators to design a series of tests tailored to different skill domains or learning objectives. The test sequence can then be randomized, providing further flexibility in assessment administration, and ensuring that each test iteration presents a unique set of challenges to the test-takers. To give an example, this feature allowed to field test and validate more than 1000 developed items for a Luxembourgish orthography test (Sonnleitner, Keller, & Sperl, 2023). An incomplete block design was established and 1000 linked but unique test versions, each encompassing 150 items were created and then implemented in and administered via OASYS. This procedure allowed for a maximum of administered items while at the same time keeping item exposure to a minimum.

While adaptive testing is not yet implemented, it is feasible with additional work in terms of methodology and IT capabilities. The primary barrier remains the quantity of items available for inclusion in the assessment pool. Without an ample supply of psychometrically validated items spanning a wide range of difficulty levels and subject domains, the implementation of adaptive testing may be hindered. Therefore, the focus must be on continually expanding and refining the item bank to ensure sufficient construct coverage and diversity, thereby unlocking the full potential of adaptive testing within the platform.

# Generative item models: experiences and the example of autoMATH

## The need of e-assessment platforms for content

One huge bottleneck for the successful application of e-assessment platforms is the (amount of) content. Although this statement might seem trivial since it is true for all kind of assessment media, such as paper-pencil, or even oral examinations, the electronic administration mode potentiates this issue by a huge factor. Phrased differently, many benefits of e-assessment, such as adaptive, branched testing, fluid tests in the case of OASYS (see above) or individualized testing, can only be leveraged if there is a vast amount of content readily available. This content - ideally being psychometrically evaluated and calibrated – should fully cover the targeted construct and span the whole difficulty range. When thinking of e-assessment platforms as elaborate databases, the need for well-curated and highly qualitative content becomes even more evident.

Content development procedures, however, have not changed one iota since the early days of large-scale assessment. This not only holds true for tech-only item formats, such as complex problem-solving scenarios (Sonnleitner, et al., 2017), but also for the "Big 3" of educational large-scale assessment: reading comprehension, mathematics, and science items. Usually, a stimulus and related items are developed by a group of subject matter experts (SEM), reviewed by other content experts and psychometricians, field tested, calibrated, and lastly included in the final item pool (e.g. Wu, Tam, & Jen, 2016). This approach is costly in terms of time and other resources and usually limits item pools being available in the test platforms.

## Model-based item development: a solution?

One attempt to solve this issue was seen in using digital devices not only for item administration but also for item generation (and therefore filling e-assessment platforms). These attempts - being subsumed under the term automatic item generation (AIG) - date back to the early 80s and usually include a sophisticated template or blueprint that is translated into computer code and used for algorithmically generating high amounts of test items (Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002). The starting point for all of these endeavours is a so-called cognitive model of what is going on in the student's mind when solving a specific task. Figure 3 gives an example of a cognitive model for the ability to calculate the sum of three numbers, a basic competency defined in the Luxembourgish school curriculum for third grade students (Basis Grades 1-2, domain Numbers and Operations). After outlining the competency to be measured, it is broken down into concrete mental operations that need to be taken into account to solve the task. Hence, different structural characteristics of the task are described that could be manipulated to generate groups of items. The example depicted in Figure 3 illustrates the potential of this model to generate diverse test items through manipulating different characteristics of the numbers used. By adjusting factors, such as decade crossing and the overall numerical range, the model can produce a wide array of variations, each differently influencing the difficulty. The incorporation of various semantic embeddings, such as adventure or sports themes, the model further expands its capacity to generate plausible test items. Consequently, the potential for generating unique and realistic scenarios becomes practically limitless, offering a rich resource for creating engaging assessments across a spectrum of topics and themes. Using such cognitive models for automatic item generation has successfully been demonstrated for a wide variety of abilities (cf. Gierl & Haladyna, 2013 or Gierl, Lay, & Tanygin, 2021).

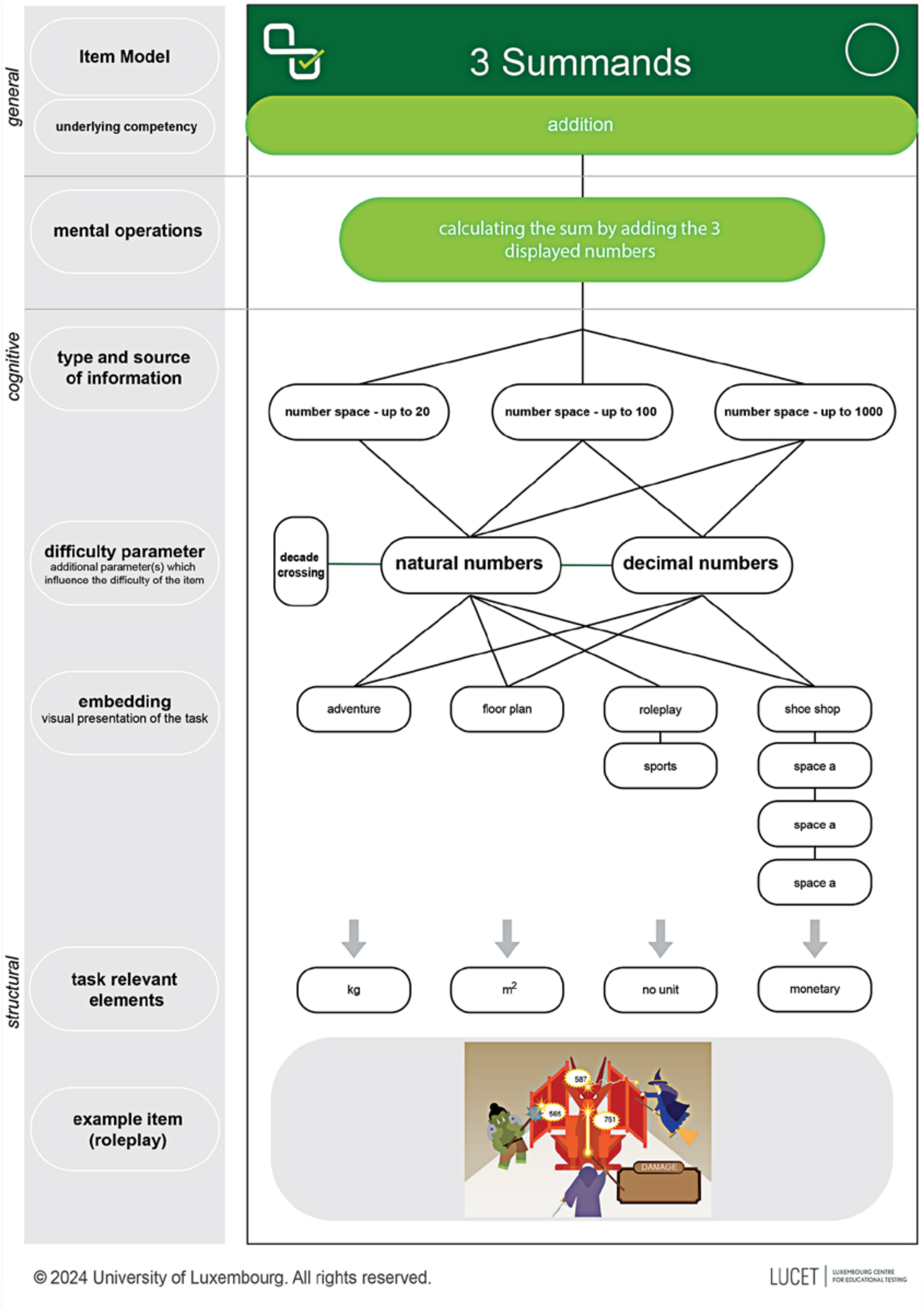


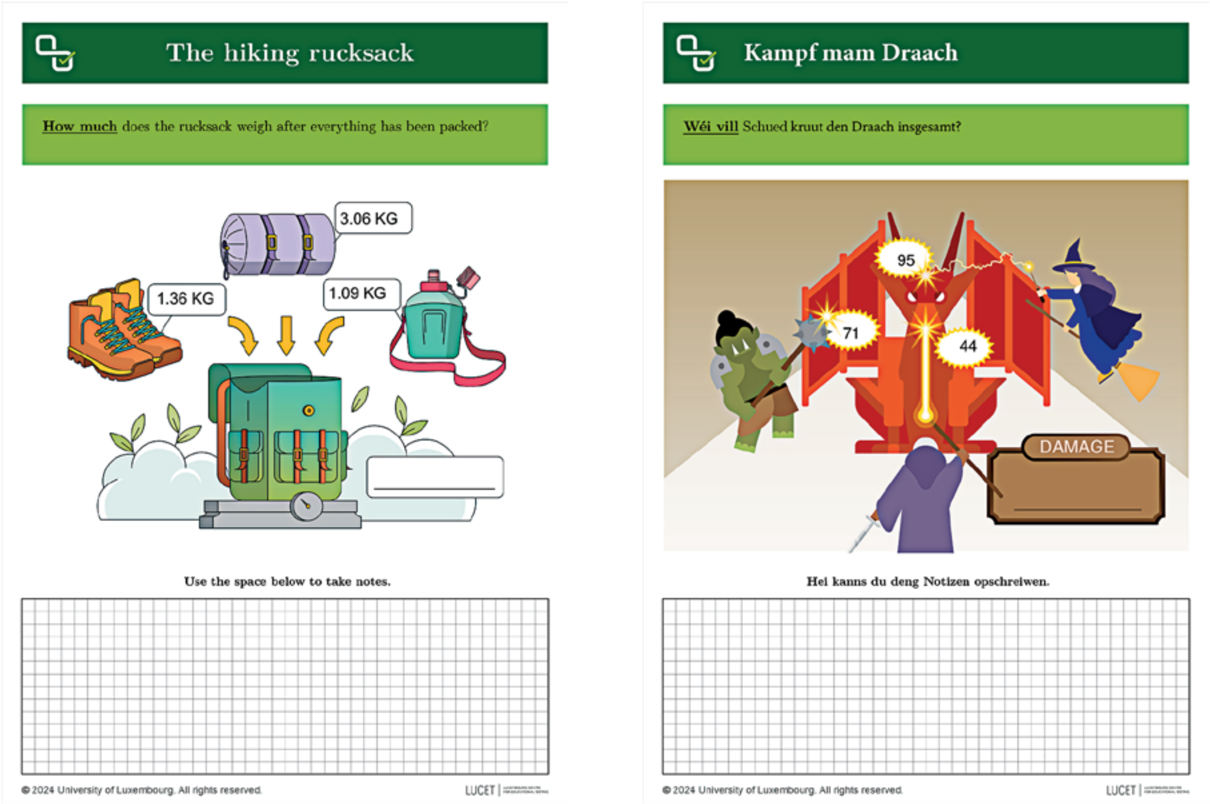
**Figure 3**: Cognitive model for the competency to calculate the sum by adding 3 numbers

Ideally, cognitive models are based on empirical findings or theoretical considerations on the underlying mental mechanisms of a certain competency. Instead of single items, SMEs develop cognitive models based on their knowledge, thus guaranteeing a high degree of content validity and control over the generated items. This approach therefore not only has the pragmatic advantage of maximizing the output of the experts' time, but it also helps to make implicit knowledge of item developers explicit and therefore tangible, reproducible, and item development itself more accountable. Whereas previous attempts to use such models to predict item difficulties have delivered mixed results, they certainly provide added value when it comes to explaining unexpected psychometric characteristics of items or could build the base for more advanced diagnostic data analysis, such as Cognitive Diagnostic Models (cf. von Davier & Lee, 2019). When it comes to competencies defined in school curricula (often an amalgam of more fine-grained abilities), elaborate cognitive theories as a basis for such models, let alone their empirical validations, are however rare (Leighton & Gierl, 2011).

## Experiences with model-based item development

Given the promising advantages of model-based item development, and to explore its potential within the Luxembourgish school monitoring, in 2020 the LUCET started a research project funded by the national research agency FNR (FAIR-ITEMS, C19/SC/13650128). In total, 55 cognitive models were developed for the mathematical domains of numbers & operations (32) and space & form (23), spanning the elementary school curriculum from Grade 1 up to Grade 6.

We adopted the standard procedure in AIG and started by identifying so-called parent items, i.e. items that perfectly represent a certain competency and proved to be psychometrically sound in previous test administrations. Those "parents" were then analysed by teachers and psychometricians to identify elements that could be manipulated and likely had an impact on item difficulty. As expected, cognitive studies helping at this stage were rare (notable exceptions being basic arithmetic competencies), so we drew on teachers' experiences or recommendations from the curriculum to identify the relevant levers for manipulation. One particular challenge that we faced was LUCET's commitment to language-reduced item formats, i.e. using mostly illustrations and only little text to provide the tasks instructions. Since language is a known predictor of mathematics performance in Luxembourg's highly heterogeneous student population (cf.  Greisen et al., 2021), using illustrations was a necessity but at the same time true pioneering work in the field of AIG that mainly dealt with text-based or graphically simple items before. Thus, we closely collaborated with a graphic agency to prepare highly structured but nevertheless appealing sets of illustrations that we could use for item generation. It is important to note that illustrations were kept simplistic to reduce students' cognitive load when working on them. After translating the cognitive model's logic into programming language (in our case, we referred to R), those graphics where then used to compile items (see Fig. 4 for example items measuring the competency of adding three numbers).



The hiking rucksack

How much does the rucksack weigh after everything has been packed?

Use the space below to take notes.

© 2024 University of Luxembourg. All rights reserved.

LUCET

Kampf mam Draach

Wéi vill Schued kruut den Draach insgesamt?

Hei kanns du deng Notizen opschreiwen.

© 2024 University of Luxembourg. All rights reserved.

LUCET

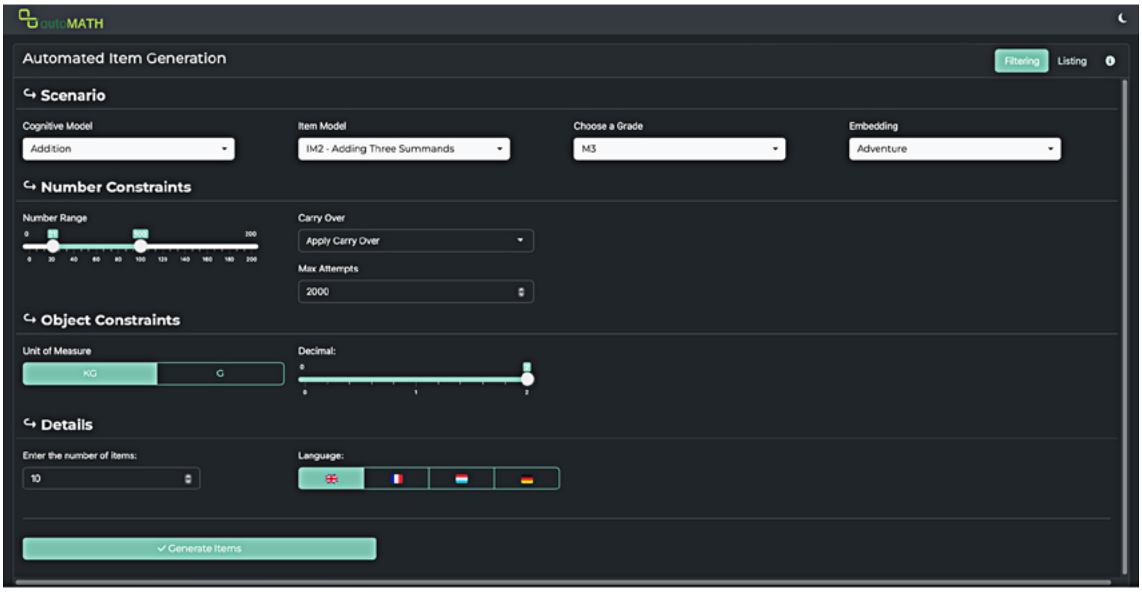
**Figure 4**: Two generated items for the competency of adding three summands including the cognitive constraints “decimal numbers” (left) and “decade crossing” (both). These examples also demonstrate the models’ language implementation in English, Luxembourgish, German, and French

Our approach was so convincing that in 2021 we studied the psychometric stability and fairness of 24 cognitive models in the domain of numbers & operations (Inostroza et al., 2023). In total, 402 items were generated (maximum 18 items per model) that were systematically varied concerning difficulty inducing components and semantic embeddings and then administered in Grades 1 (n = 2704), 3 (n = 4126), and 5 (n = 3549). Results showed that for about half of the models, psychometric difficulties of the items could be fully explained by model parameters, pointing to stable and potentially predictable item characteristics. The semantic embedding the tasks were presented in, impacted item difficulty especially in the lower grades and six models contained embeddings that caused subgroup differences. It is worth noting however, that this was mostly an issue in Grade 1 and not in Grade 5, pointing to the fact that assessment in younger students contains much more noise, especially when illustrations are used that might trigger rich associations in the children’s minds.

## A versatile item generator for elementary school mathematics: the autoMATH

Based on the promising results of the 2021 study, we decided to take model-based item generation one step further and started developing autoMATH in early 2024, an app to automatically generate elementary school math items. The existing R code was overhauled, model related information was stored in a dedicated SQL database (before, we relied on Excel sheets) and a user interface using R-shiny was programmed. Due to the template-based structure of all developed cognitive models, it was easy to add additional languages for each semantic embedding. Figure 5 presents the current user interface or front-end. After selecting the relevant competency, the user can choose the Grade or Age group the items should be generated for. Depending on the choice, certain constraints on the number range are automatically set (e.g. number range of 0-20 for Grade 1, 0-100 for Grade 3, and 0 to 1000 for Grade 5) but could be deliberately changed as well. For each model, different semantic embeddings are available which automatically impose certain constraints on the generated numbers (e.g. hiking shoes only having a certain weight range). In addition, it can be selected if the resulting addition problem contains decade crossing (i.e. the sum of digits exceeds 9 and students need to “carry over” the extra value) or decimal numbers, allowing for the generation of specific items targeted at certain competencies. After defining these elements, the number of generated items and the language of presentation can be chosen. Currently, all implemented models can generate items in English, Luxembourgish, German, and French. Adding further languages would be relatively easy since only respective columns would be needed in the underlying data base.

Currently, more cognitive models are consecutively implemented, and items are generated in pdf-format. This could be quickly changed though to all kinds of image formats depending on the specific needs of the test setting, e.g. pngs for web-based administration.



autoMATH

Automated Item Generation

Scenario

Cognitive Model: Addition

Item Model: IM2 - Adding Three Summands

Choose a Grade: M3

Embedding: Adventure

Number Constraints

Number Range: 0 – 200

Carry Over: Apply Carry Over

Max Attempts: 2000

Object Constraints

Unit of Measure: KG | G

Decimal: 0 – 2

Details

Enter the number of items: 10

Language:

**Figure 5:** User interface of the autoMATH item generator (Sonnleitner et al., 2024)

It is easy to see, how model-based item generators, such as autoMATH could be used to address the huge demand for well curated content in e-assessment or e-learning platforms. With a simple click of a button, a plethora of math items tailored to specific competency levels could effortlessly be generated, numbering in the millions. Given the results of our 2021 study, it is fair to say that models could be developed which produce valid items with predictable psychometric characteristics and negligible subgroup differences despite substantially different surface characteristics - one major requirement for item banks being used for adaptive or branched, or even fluid testing (see above). A further integration of generative models into such testing or learning platforms, enabling true on-the-fly item generation would even allow for truly individualized test or learning content that is presented in a way that is chosen by the students themselves – potentially impacting test taking motivation and commitment. Furthermore, the advantages for test security are readily apparent. The abundance of available items allows for a significant reduction in item exposure, as the sheer volume of options minimizes the likelihood of repetition. Alternatively, each item could be utilized just once, thereby further mitigating the risk of compromise, and ensuring the integrity of assessments which remains paramount.

Besides these more technical advantages, it is worth mentioning that item writing through cognitive model building helped to achieve a much better understanding of the assessed construct. Defining the models forced item writers and SMEs (teachers and item developers being trained in mathematics specifically) to rigorously analyse the targeted competency and already developed items on a very fine-grained level. Not to mention that cognitive models are transparently documenting content validity evidence of the assessed competencies, further contributing to the validity of the whole assessment.

These manifold advantages of model-based item generation, however, come at certain costs: First, the overall development required an extensive multi-disciplinary team of SMEs, psychometricians, web developers, and illustrators. In our case, most expertise was found in-house at LUCET, but it is unlikely not to rely on (costly) external expertise and agencies. Second, setting up a standard operating procedure for a) developing, and b) implementing cognitive models into a generator was quite a challenge given the complexity of the models and the interplay of various team members to develop and integrate them. Finally, model development itself poses several challenges, especially when real-world semantic embeddings are used. After establishing the right "granularity" of the model (e.g. how many competency levels should a model cover, how many different semantic embeddings should be available), it is far from trivial to decide which elements should be manipulated (as stated above, literature is scarce on these aspects) and in which way (e.g. price range of objects, weight of animals). Translating all identified constraints in computer code and appropriately preparing graphic files (e.g. defining the size and position of text boxes) are additional challenges.

## ...but what about using AI?

Since their breakthrough in late 2022 by the release of ChatGPT, generative large language models (LLMs) and Artificial Intelligence (AI) in general became the focus of public discussions, hopes, and fears due to their impressive generative capabilities. The same holds true for AI-based image generators, such as Dall-E, Stable Diffusion or Midjourney that were released roughly at the same time. Those technologies are astounding, and it did not take long since they found their way into first applications for e-assessment, and even (mostly text-based) item generation (cf. Yaneva & von Davier, 2023). When reflecting on the use of generative cognitive models for e-assessment and e-learning, immediately the question arises, whether the approach presented above is not merely beating a dead horse. Although given the breath-taking speed of development in this field, still many questions have to be answered before item generation could be substantially assisted, let alone be fully carried out by AI. Looking at Gartner's technology hype cycle (see above), with generative AI clearly being around its peak of inflated expectations according to the tech consultancy (Gartner, 2023), we clearly must prepare for the trough of disillusionment by discovering challenges. Despite general concerns mainly tackling ethical or sustainability aspects, the European Union's "AI Report by the European Digital Education Hub's Squad on Artificial Intelligence in Education" (Le Borgne et al., 2024) identifies the following challenges that we deem highly relevant in this regard and for which we still see advantages of (conventional) model-based item development:

**Unclear ownership:** Omnipresent is the question of ownership rights over AI-generated content raising concerns regarding the allocation of intellectual property rights. Determining whether creators of the AI, developers who trained it, or users who enter the data hold the rights is crucial for establishing legal and ethical frameworks. The provenance of information used by AI systems presents challenges related to data quality, bias, and reliability. Understanding and knowing the sources of this information would be imperative to ensure its accuracy and integrity. In other words, even if the generated content would be perfectly suited, (at the moment) there is a question mark whether using this content would be copyright infringement. It is important to note that this question might be of special relevance to EU countries given their (sometimes) stricter legislation concerning intellectual property rights. A solution to this would be the training of generative AI on own/ creative common licensed text corpora or image collections; whether this is feasible for educational research institutes or testing companies is a different question though.

**Content (in)consistency:** Currently, the stability and robustness of generated content is an additional question mark. Since the generative process is opaque, it is not predictable what kind of content is created and whether this content fulfills certain quality criteria, e.g. phenomena of "hallucination" exist where generative AI produces incorrect or misleading results. See Figure 6 for two examples using Dall-E/ChatGPT4 for so-called zero-shot (no previous training) generation of images similar to those used in the cognitive model presented above. Although visually quite appealing, it becomes evident that it would require further attempts to get usable content. Although results can certainly be improved by careful and precise "prompt engineering" (i.e. the request that is given to the AI algorithm, e.g. Sayin & Gierl, 2024), refining this process would take time and nevertheless require a final, manual check of the generated content. This, in turn, would undermine the very purpose of using AI to automate tasks: such an inspection process, would be extremely time-consuming and cost-inefficient, negating the benefits of AI-driven automation.

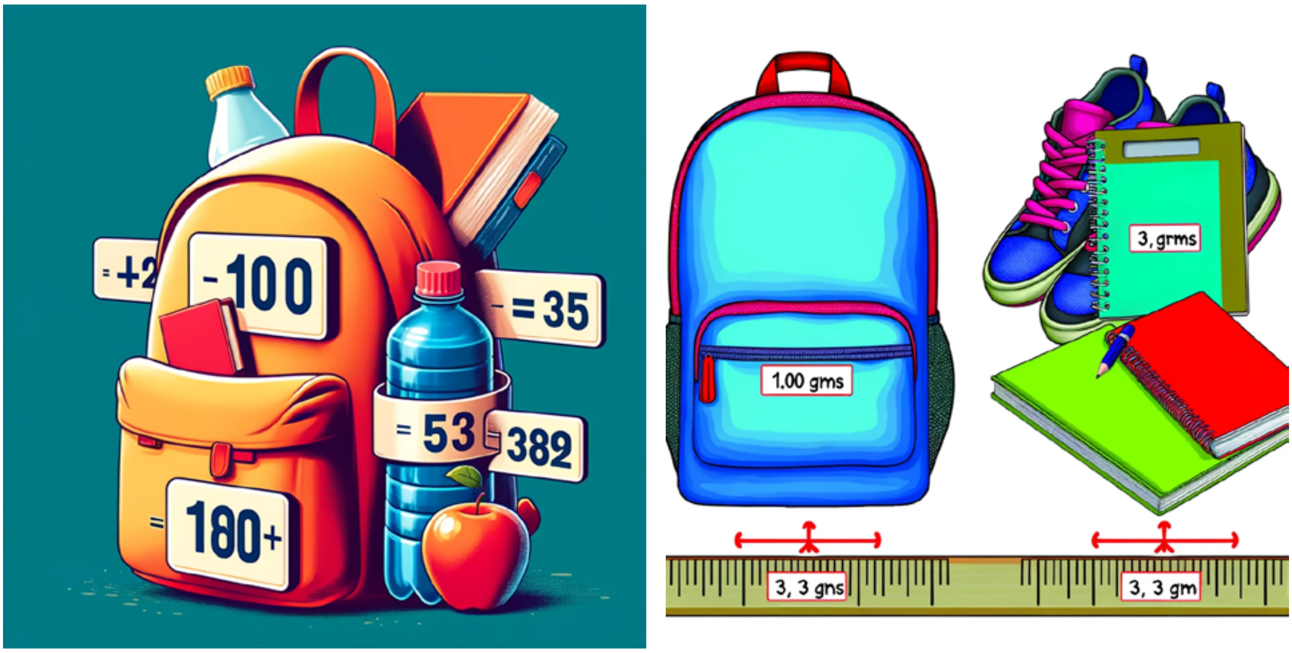


**Figure 6:** first two images generated by Dall-E/ChatGPT4 using the prompt "Draw me an image of a backpack. This backpack is being packed with three objects. Each object has a weight tag to it in grams. The total weight of the three objects should not exceed 1000g."

**Intransparency of generative process:** Due to their highly complex nature, generative algorithms, such as LLMs are hardly understood and therefore often called blackbox-systems. The Cornerstones of educational and psychological assessment, such as validity or the use of unbiased content (cf. the Standards for Educational and Psychological Testing, AERA, APA, & NCME, 2014), all require full traceability of item writing decisions. Making the prompts usable for generating content transparently builds a first step, but the more decisions (e.g. how competency levels are defined) are handed over to the AI, the more opaque and therefore problematic it becomes.

Clearly, for these aspects (ownership, content consistency, and transparency) solutions need to be (and will be) found. Model-based approaches as presented for example in the autoMATH above, however, provide full controllability from the outset and therefore ensure full accountability – a key aspect in educational settings.

## Discussion & Conclusion

In this article, we have explored two essential components of e-assessment platforms by looking at case studies of the Luxembourgish Centre for Educational Testing: A platform's form and design by using the Luxembourg originated platform OASYS as example, and a valid and scalable way to create content by using the autoMATH item generator. Combining these elements promises to unfold the full potential of e-assessments as already foreseen during its rise. By utilizing technology to streamline the assessment process and adopting a model-based approach to item development, we can enhance the quality, validity, and efficiency of assessments. This combination allows for greater customization and adaptability in assessment design, ensuring that assessments accurately measure the desired constructs while minimizing biases and errors. In addition, it opens up venues to increase students' engagement with the tests through the possibility of customization.

While Artificial Intelligence (AI) holds enormous potential for revolutionizing e-assessment practices, we currently see too many open questions related to intellectual property, inconsistent content, and intransparency of the generative process. By opting for a model-based approach, we maintain control over the content creation process, ensuring consistency, reliability, and transparency in assessment design.

However, our experiences have revealed that achieving our goals is easier said than done, as significant investments are required to develop and implement such an advanced e-assessment platform and model-based item generator. Despite these challenges, the Luxembourg Centre for Educational Testing (LUCET) remains committed to this endeavour, recognizing it as an investment in the future of educational assessment with the humble hope of providing an example and inspiring other institutions in this field.

## Acknowledgements

The authors want to specifically thank Katarina Aleksić for reviewing and commenting an earlier version of the chapter, and Cecile Braun, Eric François, Ulrich Keller, Yannick Rathmacher, and Pierrick Kinif for their support.

## References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington: American Educational Research Association.

Blosch, M. & Fenn, J., Understanding Gartner's Hype Cycles, Gartner, 2018. Retrieved 16/05/24 from: https://www.gartner.com/en/documents/3887767

Bunderson, C . V., Inouye, D. K . & Olsen, J. B. (1988). The four generations of computerized educational measurement [ETS Research Report]. ETS.

Chandrasekaran, A., and Davis, M. (2023). Gartner, Hype Cycle for Emerging Technologies. Gartner. Retrieved 16/05/24 from: https://www.gartner.com/en/documents/4597499

Fischbach, A., Greiff, S., Cardoso-Leite, P., & Koenig, V. (2021). Digitalisierung der pädagogischen Diagnostik: Von Evolution zu Revolution. Nationaler Bildungsbericht Luxemburg 2021, 136-140.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). Automatic item generation: Theory and practice. Routledge.

Gierl, M. J., Lai, H., & Tanygin, V. (2021). Advanced methods in automatic item generation. Routledge.

Greisen, M., Georges, C., Hornung, C., Sonnleitner, P., & Schiltz, C. (2021). Learning mathematics with shackles: How lower reading comprehension in the language of mathematics instruction accounts for lower mathematics achievement in speakers of different home languages. Acta Psychologica, 221, 103456.

Holmes, W., Persson, J., Chounta, I. A., Wasson, B., & Dimitrova, V. (2022). Artificial intelligence and education: A critical view through the lens of human rights, democracy and the rule of law. Council of Europe. https://rm.coe.int/artificial-intelligence-and-education-a-critical-view-through-the-lens/1680a886bd

Inostroza Fernandez, P. I., Michels, M. A., Hornung, C., Gamo, S., Keller, U., Gierl, M., Cardoso-Leite, P., Fischbach, A., & Sonnleitner, P. (14 April 2023). *The impact of cognitive characteristics and image-based semantic embeddings on item difficulty* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education.

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2013). Item generation for test development. Mahwah, NJ: Erlbaum.

Le Borgne, Y.-A., Bellas, F., Cassidy, D., Vourikari, R., Kralj, L., Obae, C., Pasichnyk, O., Bevek, P., Deyzen, B. van ., Laitala, A., Sharma, M., Wulgaert, R., Niewint-Gori, J., Gröpler, J., Joyce, A., Rondin, E., Gilleran, A., Janakievska, G., Weber, M., & E. D. E. H. S. on A. I. in Education. (2024). AI report (Version 1). Royal College of Surgeons in Ireland.

Leighton, J. P., & Gierl, M. J. (2011). The learning sciences in educational assessment: The role of cognitive models. Cambridge University Press.

Miao, F., & Holmes, W. (2023). Guidance for generative AI in education and research. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000386693

Organisation for Economic Co-operation and Development (2016). PISA 2015 technical report. Paris: OECD.

Rohles, B., & Backes, S. (2021). Wissen zu Nachhaltigkeit und Verständnis für komplexe Zusammenhänge. Eine Concept-Mapping-Studie. In SCRIPT & LUCET (Ed.), Nationaler Bildungsbericht Luxemburg 2021 (pp. 160-166). Esch-sur-Alzette, Luxembourg: University of Luxembourg.

Rohles, B., Koenig, V., Fischbach, A. & Amadieu, F. (2019). Experience matters: Bridging the gap between experience- and functionality-driven design in technology-enhanced learning. Interaction Design and Architecture(s) Journal, 42, 11–28.

Sayin, A., & Gierl, M. (2024). Using OpenAI GPT to Generate Reading Comprehension Items. Educational Measurement: Issues and Practice, 43: 5-18.

Sonnleitner, P. (2019). Gamification of psychological tests: three lessons learned. Testing International, 42.

Sonnleitner, P., Keller, U., Martin, R., Latour, T., & Brunner, M. (2017). Assessing complex problem solving in the classroom: Meeting challenges and opportunities. In B. Csapó & J. Funke (Eds.), The nature of problem solving (pp. 169–187). Paris, France: OECD.

Sonnleitner, P., Keller, U., & Sperl, H. (2023). Entwicklung und Validierung des Luxemburger Orthografietests. Luxembourg Centre for Educational Testing, University of Luxembourg.

Sonnleitner, P., Kinif, P., Bernard, S., & Rathmacher, Y. (2024). autoMATH – automatic math item generator. [Computer software]. Luxembourg Centre for Educational Testing, University of Luxembourg.

von Davier, M., & Lee, Y.-S. (2019). Handbook of diagnostic classification models: Models and model extensions, applications, software packages. New York: Springer.

Wu, M., Tam, H. P., & Jen, T.-H. (2016). Educational measurement for applied researchers: Theory into practice. Singapore: Springer.

Yaneva, V., & von Davier, M. (2023). Advancing natural language processing in educational assessment. NCME educational measurement and assessment book series. Taylor & Francis.

# Netherlands

# NETHERLANDS: BIOGRAPHIES

## Max van der Velde

Max van der Velde is a PhD candidate of the Cognition, Data and Education section of the Behavioural, Management and Social sciences (BMS) faculty of the University of Twente. His main research interests include reading fluency, research methodology, and educational assessment.

## Bernard Veldkamp

Prof.dr.ir. Bernard Veldkamp is Vice-Dean of Research of the Faculty of Behavioral, Management and Social Sciences (BMS). His main research interests include methodology, measurement and data analytics within the context of educational, psychological, and health sciences.

## Jos Keuning

Dr. Jos Keuning is the head of the Educational Research department of CitoLab at Cito. His main research interests include educational assessment and reading competency.

## Remco Feskens

Dr. Remco Feskens is director of CitoLab at Cito, and a research fellow at Cognition, Data and Education at the Faculty Behavioural, Management and Social sciences (BMS), University of Twente. His main research interests include educational assessment and psychometrics.

## Nicole Swart

Dr. Nicole Swart is a senior researchers at the Dutch Center for Language Education [*Expertisecentrum Nederlands*] where she is mainly involved in large-scale assessment studies, both international and national (i.e. PIRLS, IELS, Peil. Onderwijs). Her main interests include reading fluency, reading comprehension and early childhood development.

## Wieke Harmsen

Wieke Harmsen is a PhD candidate at the Centre for Language Studies of the Faculty of Arts at Radboud University in Nijmegen. In her research, she uses language and speech technology to automatically measure children's reading and spelling skills.

# THE FRAMEWORK AND DEVELOPMENT OF SERDA: SPEECH ENABLED READING FLUENCY ASSESSMENT FOR DUTCH

## Abstract

The importance of reading for educational, vocational and societal life cannot be understated. Nonetheless, recent large-scale studies reveal that the reading comprehension of students has declined globally, and specifically in the Netherlands. Developing fluent reading skills allows children to read quickly, accurately and with proper expression, which is fundamental to become a good reader. To monitor this development, teachers need to assess fluency on a regular basis. However, fluency assessment is currently time-consuming for teachers, provides limited information, and neglects prosody assessment. This chapter presents a framework for, and the development of, a digital automatic fluency assessment tool for early primary education that overcomes current issues through incorporating Automatic Speech Recognition (ASR): the Speech Enabled Reading Diagnostics App (SERDA). Three word- and passage reading tasks were developed based on popular pen-and-paper instruments, and administered to 653 primary school children. The results provide usability, validity and reliability evidence for SERDA's speed and accuracy measures. Furthermore, SERDA reduces the testing burden placed on teachers, increases the information gained, and facilitates prosody assessment.

Wordcount: 5483

## Introduction

Being a proficient reader is essential to succeed throughout educational, vocational, and societal life (Horning, 2007). Nonetheless, recent large-scale studies reveal that the reading ability of fourth grade students has been on the decline globally (Mullis et al., 2023), and especially in the Netherlands (Swart et al., 2023). In addition, while less than a fourth of Dutch fifteen year-olds were found to be at risk of functional illiteracy in 2018 (Gubbels et al., 2019), recent research shows this now concerns every third (Meelissen et al., 2023). A means to counteract this trend is found through improving the development of fluent reading skills, a widely acknowledged and critical component for the development of proficient reading (National Institute of Child Health and Human Development, 2000). Given that the monitoring of this development requires teachers to assess the fluency of children's reading at a regular basis, and given that current assessment practices know many shortcomings, improving the assessment of reading fluency could help impede the imminent increase in functional illiteracy.

Reading fluency is defined as the ability to read quickly, accurately and with proper expression (Kuhn et al., 2010; Pikulski & Chard, 2005). The speed and accuracy of reading are often referred to as the automaticity of reading (e.g. Kim et al., 2021). This conceptualization dates back to the rationale discussed by Logan (1988), who argues that once a person has had sufficient practice, allowing them to read both quickly and accurately, reading becomes automatic. Here, automaticity indicates that reading requires little effort, which frees up cognitive resources, and allows the reader to focus on more complex aspects of reading, such as comprehending (Aldhanhani & Abu-Ayyash, 2020; Morris & Perney, 2018). The remaining component, expressiveness or prosody, is described by the literature as the ability to properly use a combination of phrasing, expression, intonation, stress, pitch, and pauses (van der Velde et al., 2024). The ability to read expressively has previously been linked to both earlier and later reading comprehension, the directionality of the relationship being dependent on children's primary school Grade (Veenendaal et al., 2016). To summarize, reading fluency can be seen as the degree of automaticity, or speed and accuracy, and prosody of reading.

While the construct of reading fluency is generally agreed upon, its assessment has proven problematic. Fluency assessment currently focusses on the number of words read correctly per minute (WCPM), which is an operationalization of automaticity rather than fluency (Benjamin et al., 2013; van der Velde et al., 2024). This focus on WCPM is found for popular international instruments such as the Dynamic Indicators of Basic Early Literacy Skills Oral Reading Fluency (DORF; University of Oregon, 2020), as well as for popular instruments in the Netherlands for both word reading (i.e., the Three Minute Task [Drie-minuten-toets; DMT] (van Til et al. 2018a)) and passage reading (i.e., AVI [Analyse van Individualiseringsvormen; AVI] (van Til et al. 2018b)). Though this underrepresentation of prosody assessment is nothing new (Paige et al., 2017), it persistently influences the validity and viability of using fluency scores in practice.

That is not to say that the overrepresentation of automaticity assessment is incomprehensible from a practical point of view. Automaticity assessment can generally be conducted both swiftly and easily (e.g. University of Oregon, 2020). Meanwhile, prosody assessment tends to be more complicated, as it requires further training, demands the administration of a separate instrument, and provides relatively subjective information (Kuhn et al., 2010). Given that test administration and scoring is work that is mostly carried out by teachers, placing a heavy testing burden on them, it is not difficult to understand why incorporating prosody assessment is often deemed too time-consuming. Moreover, even when only taking into account automaticity assessment, the extraction of detailed diagnostics tends to be limited in practice, as these require a more thorough and time-consuming investigation of the reading performance.

In short, the assessment of reading fluency is overrepresented by its speed and accuracy components. In addition, fluency assessment currently places a large testing burden on teachers and does not yield detailed diagnostics when conducted in a practically feasible manner. Therefore, creating an assessment tool that could limit teacher burden while providing detailed and objective fluency diagnostics on all fluency components could considerably help teachers, children and society at large. In this chapter, we describe the proposed framework to overcome these assessment shortcomings, which will subsequently be implemented within the development of a reading fluency assessment instrument referred to as the Speech Enabled Reading Diagnostics App (SERDA).

## SERDA's Framework

SERDA's framework describes how to improve reading education at the primary school level by means of assessing reading fluency through the analysis of speech from reading aloud tasks, and by means of modelling the resulting data to provide individualized feedback on how to improve reading. Specifically, the final goal is for SERDA to visualize the reading ability of children for teachers at both the class and individual level. Information should be presented on children's general ability to read fluently, as well as more specific information on the speed, accuracy and expressiveness of reading. In addition, SERDA should be able to differentiate between children's performance on the reading of word lists and passages, providing a comparison of proficiency in context-free and context-specific reading.

In order to manifest these ambitions, SERDA's framework combines automatic speech recognition (ASR), speech diagnostics and learning analytics to create an innovative, integrated approach to reading diagnostics and automated feedback, as illustrated in Figure 1. Throughout this framework, ASR concerns the "independent, machine-based process of decoding and transcribing oral speech" (Levis & Suvorov, 2012, p. 1). Speech diagnostics refer to the relevant speed, accuracy and expressiveness measures extracted from speech data by the ASR-algorithm. Learning analytics is generally defined as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Society for Learning Analytics, 2011). Within the context of this framework, learning analytics specifically relate to the analyses conducted to transform speech diagnostics into recommendations that can be used to improve personal learning-to-read trajectories in primary education. In essence, when compared to its pen-and-paper contemporaries, SERDA's most fundamental difference is its usage of speech data and the transformation of speech into relevant diagnostics through ASR.

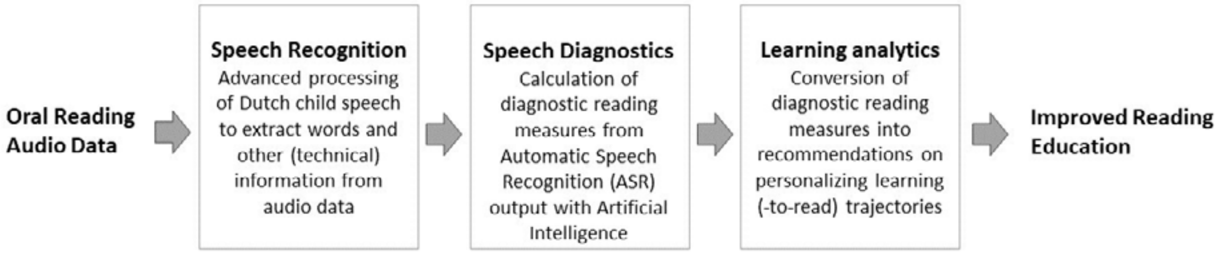

Oral Reading Audio Data
Speech Recognition
Advanced processing of Dutch child speech to extract words and other (technical) information from audio data
Speech Diagnostics
Calculation of diagnostic reading measures from Automatic Speech Recognition (ASR) output with Artificial Intelligence
Learning analytics
Conversion of diagnostic reading measures into recommendations on personalizing learning (-to-read) trajectories
Improved Reading Education

**Figure 1** SERDA's Reading Fluency Assessment Framework

### Automatic Reading Fluency Assessment

Research on implementing ASR with the goal of improving reading ability dates back more than two decades (Mostow et al., 2003; Reeder et al., 2007). This incorporation was fruitful, as evidenced by the wealth of successful research on ASR-based reading tutors for reading practice and automatic assessment in English (Bolaños et al., 2013; Loukina et al., 2019; Sabu & Rao, 2018) and other languages (Godde et al., 2017; Proença et al, 2015; Silva et al., 2021).

Within the context of the Dutch language, Bai et al (2020) have recently shown that ASR can be successfully used to assess and provide feedback on the reading accuracy and speed of first graders. In addition, Wei et al (2022) showed the potential of ASR in assessing reading errors for non-native speech. Indeed, up to now, much research on the use of ASR for the speech of Dutch children has focussed on extracting reading errors (Nicolao et al., 2018; Yilmaz et al., 2014), while much less work has focussed on prosody (e.g. Cucchiarini et al., 2000). However, previous research has shown that it is possible to extract automatic measures from speech that are related to subjective fluency and prosody ratings (Benjamin et al., 2013; Cheng, 2011; Chung & Bidelman, 2022; Dimzon & Pascual, 2023; Truong et al., 2018).

### The present study

Reading fluency assessment in the Netherlands currently places too large a testing burden on teachers and does not yield detailed diagnostics when conducted in a practically feasible manner. The current study presents a framework to overcome these shortcomings. Based on this framework a reading fluency assessment instrument was developed. Throughout the remainder of this chapter, we will discuss the development of this instrument and the collected speech data. In addition, we provide usability, reliability and validity evidence to substantiate the use of SERDA's speed and accuracy measures in practice. The extraction and evaluation of SERDA's prosody measures is discussed in another paper, as these require different algorithms and methodology, which reaches beyond the scope of this chapter.

## Methods

The development of SERDA followed the following steps: First, we constructed reading tasks in collaboration with subject-area experts based on currently popular fluency instruments. Then, we administered the reading tasks to 653 children attending Grade 2 and 3 of primary schools in the Netherlands. In addition, to evaluate the validity of SERDA's tasks we obtained the most recent results of the DMT and AVI, which are the standard Dutch fluency instruments used throughout primary education.

## Task-Development

### SERDA: Word reading

In order to assess the ability of children to read context-free words, three Dutch word lists were developed based on the DMT (Van Til, 2018a). Each list contained 50 words. The first set consisted of one-syllable words with varying consonant-vowel (cv) combinations, cv/vc/cvc/ccv/ccvc/vcc/cvcc/ccvcc, and included various reading difficulties (i.e., *sch*-, -*ng/nk*, open syllable). The second set consisted of one-, two- and three-syllable words, including various advanced reading difficulties (i.e., *be-/ge-/ver-, -lijk*). The third set consisted of two-, three-, and four-syllable words, including various complex reading difficulties (i.e., loanwords, -*isch, -x-, -y-*). Words were chosen by experts in the field, based on a Dutch reading fluency test (Keuning & Verhoeven, 2005).

In order to obtain accurate word reading speed estimates, a progressive demasking design was used (Grainger & Sugui, 1990). During the progressive demasking task, a word was individually presented in the middle of the screen with a mask placed over the word, resulting in a seemingly empty screen. Then, the mask was removed for 17 milliseconds (ms). This left children with 17 ms to read the word, after which the mask returned and the first cycle was completed. The removal time of the mask gradually increased to 340 ms, in steps of 17ms per cycle. Children were instructed to tap the screen as soon as they recognized the word, after which they read the word out loud. When children were not able to read the word with a presentation time of 2,200 ms, corresponding to the 20th cycle, the child moved on towards the next word. An example of the masking-design is visualized in Figure 2.

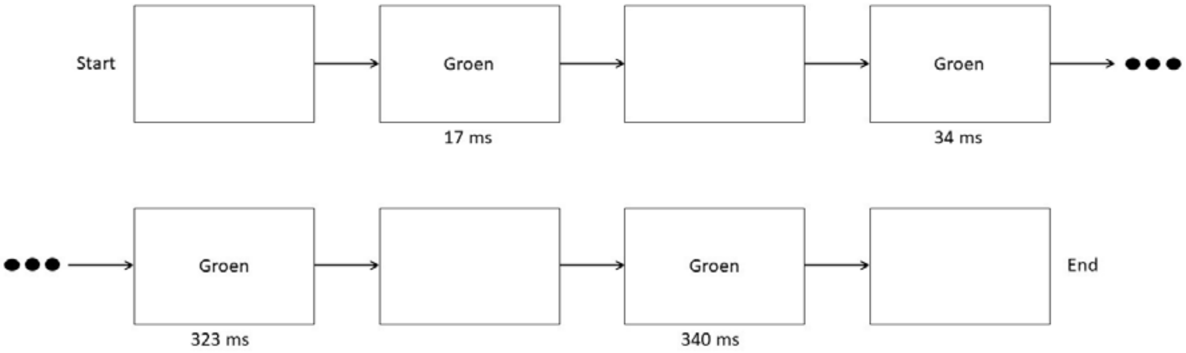


Figure 2 Example of the Start and end of a Single Progressive Demasking Trial Using the Word "Green" [Groen]

### SERDA: Passage reading

In order to assess the ability of children to read context-specific passages, three short passages of increasing difficulty were created based on guidelines that were also used in the construction of the passages of the AVI (Van Til, 2018b). Passages contained around 175 words, were written by professional child literature authors, and were selected by experts in the field. Passage-content was selected with the aim to match the interests of young children. The passages were created to, respectively, match characteristics representative of the expected reading level at the end of Grade 2, the middle of Grade 3 and the end of Grade 3. The texts contained monosyllabic, bisyllabic and polysyllabic words with incidental orthographic inconsistencies and complexities (see Van Til et al., 2018b).

## Participants

Data was collected from October 2022 to February 2023, and from October 2023 to February 2024. To establish a generalizable sample with respect to important background variables, the distribution of school-weight (Inspectie van het Onderwijs, 2024), which is an indication of children's socio-economic background, and the distribution over dialectical regions in the Netherlands (Cucchiarini et al., 2008), were used to draw two samples over all primary schools in the Netherlands. Schools for special education and special needs education were excluded from the sample. The first sample consisted of 20 school, of which 7 (35%) participated. The second sample consisted of 30 school, of which 12 (40%) participated. For participating schools, all available Grade 2 and 3 classes were included in the sample. Ethical approval was obtained from the local ethics committee, after which the approval for the participation of individual children was acquired from the children's parents or caregivers (from here: parents). The participating children were all asked to complete all reading tasks.

In total, 19 schools participated, providing a sample of 653 children. Out of all participating students 48% (n = 311) attended Grade 2, 47% (n = 310) attended Grade 3 and 5% (n = 32) attended a combined Grade 2/3 class. Boys made up 48% (n = 312) of the sample. On average, children were seven years of age when attending Grade 2 and eight years and two months of age when attending Grade 3.

The sample showed a representative distribution of schools in the Netherlands. However, it has to be remarked that the sample overrepresented schools with a low school-weight and underrepresented schools from the west of the Netherlands to some degree. With regard to school-weight, this means that relatively well-performing schools with less complex or diverse populations were more willing to participate within the current study. This finding is unsurprising, given that data collection was started a few months after the final COVID-19 lockdown was concluded. As a result, most schools in the Netherlands, and especially those with a complex or diverse population of children, were exceptionally busy and less receptive to participate in research. As for dialect region, although schools from the west of the Netherlands were underrepresented compared to the population, this resulted in a sample that provided a more equally distributed representation of schools from all dialect regions. Given the intended use of SERDA throughout all of the Netherlands, speech from children with all types of accents and dialects should be equally eligible. As a result, this underrepresentation is deemed unproblematic.

## Materials

### SERDA: Automatic Measures of Word and Passage Reading

For each reading task, SERDA yields an audio recording and a log file. The audio recordings contain the recorded speech of the child during the task. The log file contains metadata, information about the children's interactions with the application, and the total duration of the tasks. Using the audio and log data, SERDA generated item-, task- and person-level accuracy and speed measures for the word- and passage reading tasks, as well as task- and person-level WCPM-scores. An overview of all extracted metrics is presented in Table 1.

| Measure | Word-reading | Passage-reading |
|---|---|---|
| Item level | | |
| Accuracy | 0 or 1 | 0 or 1 |
| Speed | Flashing time (seconds) | Speaking duration (seconds) |
| WCPM | - | - |
| Task level | | |
| Accuracy | Number of words read correctly | Number of words read correctly |
| Speed | Words read divided by total flashing time | Words read divided by task duration |
| WCPM | Accuracy divided by total flashing time | Accuracy divided by task duration |
| Person level | | |
| Accuracy | Average task-level accuracy | Average task-level accuracy |
| Speed | Average task-level speed | Average task-level speed |
| WCPM | Average task-level WCPM | Average task-level WCPM |

Table 1 Item, Task and Passage Level Measures Extracted by SERDA.

SERDA automatically computed children's word- and passage reading accuracy based on the audio recordings. The transformation from audio to accuracy followed three steps, which are visualized in Figure 3. Firstly, each audio recording was automatically transcribed using the Dutch Large-v2 ASR model Whisper Timestamped (Louradour, 2023). This model is based on OpenAI's Whisper ASR models (Radford et al., 2023) and uses Dynamic Time Warping (DTW; Giorgino, 2009) to predict word segment timestamps. Secondly, prompts (i.e., the text of the original word list or passage) were aligned with the ASR transcription of the audio using the reversed-ADAPT algorithm for grapheme alignment (Bai et al., 2021; Elffers et al., 2013). Thirdly, each prompt word that is aligned with exactly that word from the ASR transcription is labeled as read correctly (1), while all others were labeled as read incorrectly (0), resulting in item-level accuracy scores. Followingly, the number of words read correctly were calculated for each word- and passage reading task to obtain task-level accuracy scores. Finnaly, averaging over task-level measures resulted in person-level accuracy scores for the word- and passage reading tasks.

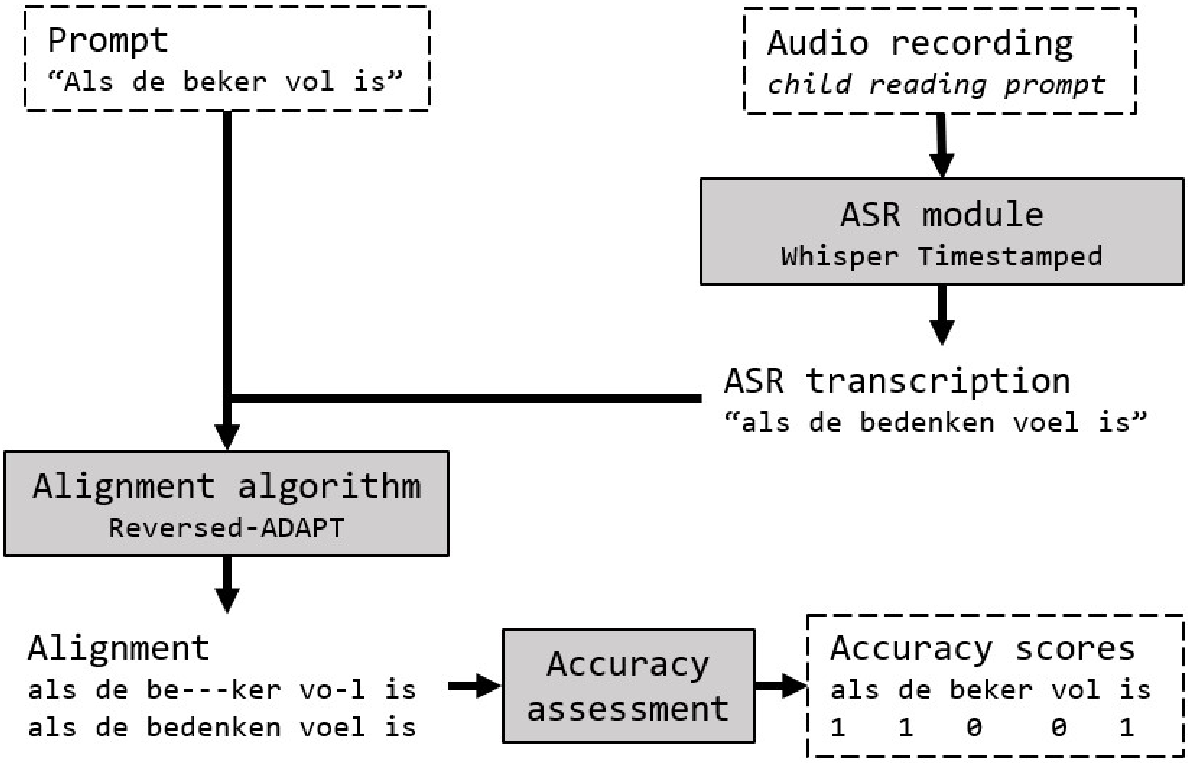


Figure 3 Visualization for the Transformation of Audio Into Accuracy Scores

The calculation of reading speed differed between the word- and passage reading task. For the word-reading task logged timestamps were used. Specifically, for each word in each word-reading subtask, SERDA stored three timestamps. The first was the onset of word presentation (T1), while the second timestamp concerned the first tap on the screen, indicating that the child has recognized the word (T2), and the final timestamp was the second tap on the screen, which signaled that the child has read the word out loud and wants to continue to the next word (T3). Then, the duration between T1 and T2, referred to as the flashing time, was calculated and functioned as the item-level speed measure. Next, the number of words read per minute (WPM) was calculated for each word-reading task by dividing the total number of words read by the total flashing time, resulting in task-level speed measures. Calculating the average WPM over all word-reading tasks resulted in person-level speed measures for the word reading task.

For the passage reading task, SERDA used the ASR output to obtain item-level speed measures. This output contains the word segments automatically recognized in the audio, together with their start and end timestamp. From the alignment of the recognized word segments with the prompt, we extract the begin and end timestamps, and thus duration, of the correctly read words. Subsequently, the task-level speed measures (WPM) were obtained by dividing the total number of words read by the total task duration. The person-level speed measures were defined as the average of the task-level speed measures.

Lastly, WCPM-scores were calculated for each word- and passage reading task. For the word-reading task, WCPM was defined as the number of words read correctly, divided by the total flashing time. For the passage reading task, WCPM concerned the number of words read correctly, divided by the time it took to complete the passage. Then, average WCPM-scores were calculated over all administered tasks, which were used as person-level measures for the word- and passage reading task.

### Word and Passage Decoding: the Three Minute Task [Drie-minuten-toets; DMT] and AVI [Analyse van Individualiseringsvormen; AVI]

The DMT and AVI are paper based tests, intended to assess the development of the word- and passage reading ability of children attending primary education in the Netherlands (van Til et al., 2018a, 2018b). Their usability, reliability and validity have been thoroughly investigated, and positively evaluated by the Dutch Committee on Tests and Testing (COTAN; Egberink & Leng, 2024a, 2024b). During the DMT and AVI, children respectively read up to three word lists or passages of increasing difficulty. Children were explicitly instructed to read the words and passages out loud as quickly and accurately as possible.

For the DMT, children had one minute to read each individual word list. The first list only contained single-syllable words with one consonant at a time. The second list also contained words with two syllables, and included words with multiple consonants. Finally, the third list allowed for the inclusion of words of more than two syllables. For the AVI, the child kept on reading increasingly difficult Grade-level passages, until the child made too many mistakes, started reading too slow, or showed a combination of both.

Then, the children's DMT and AVI classifications were obtained, which are based on their performance compared to Grade-specific Dutch norms. For the DMT, classifications concerned placing children into one of five categories, ranging from the 20% lowest- to the 20% best performing children. For the AVI, children were classified into categories that correspond to, and represent, the Grades of primary education in the Netherlands.

## Procedure

SERDA's tasks were individually administered by a test-leader in a quiet room at the children's schools, without any additional personnel. Participants conducted the tasks on a Samsung Galaxy Tab A6 tablet, while their speech was captured using a headset with an in-build microphone. To mimic current assessment practices, administration was conducted during school-hours. With the exception of school- specific break timings and opening hours, no variation existed with regard to the timing of assessment between schools. In order to evaluate the children's experience with SERDA, as well as practical and technical issues, test-leaders recorded noteworthy situations and comments from children.

During the administration of SERDA's reading tasks, children were first introduced to the word reading task. Due to the novel nature of the task, children were first taught how to correctly conduct the exercise by completing three examples with the test-leader. Once children felt comfortable with the word-reading task they completed the first word list, which contained the easiest words. After completing the first word-list the second and third word-reading tasks were conducted.

The passage-reading task was very recognizable to the children, as most have had experience with the AVI. Therefore, instructions were limited. For each passage, students were instructed to read the passage out loud, including the title, as quickly and accurately as possible. A passage was completed once the child read the entire passage, or after three minutes had passed. After the instructions, the child was allowed to start on the first passage, followed by the second and third passage.

The AVI and DMT measures were provided by the schools of the children, all of whom were familiar with conducting and scoring the AVI and DMT.

## Data analysis

Data analysis was primarily aimed at obtaining an indication of the usability, reliability and validity of SERDA's reading tasks. The usability of SERDA's reading tasks was investigated by comparing their average administration duration to those of the DMT and AVI. For SERDA's word-and passage reading tasks, this concerns the duration of administering all subtasks. For the DMT and AVI, we took the sum of the administration and scoring duration, as reported by COTAN (Egberink & Leng, 2024a, 2024b). Scoring-time was excluded for SERDA's reading tasks, as SERDA performs this task automatically.

To evaluate the reliability of SERDA's reading tasks we evaluated the internal consistency and split-half reliability. The internal consistency was evaluated by calculating Cronbach's Alpha (Cronbach, 1951) for the accuracy-scores of all items in the word- and passage reading tasks, as well as for their separate subtasks. Split-half reliability was estimated for the item-level accuracy, speed and WCPM measures of the word- and passage reading tasks by correlating 10.000 randomly generated 50/50 splits.

Then, to evaluate the validity of SERDA's reading tasks, we investigated their construct validity. Construct validity, which concerns the degree to which a tool measures the construct it aims to measure (American Educational Research Association [AERA] et al., 2014; Reynolds & Livingston, 2021), was determined by correlating the WCPM-scores of SERDA's word- and passage reading tasks with one-another. In addition, we used Spearman's Rho to compare SERDA's WCPM scores to the DMT and AVI classifications. All analyses were conducted in RStudio (version 4.3.1; Posit Team, 2023).

Finally, the feedback of test leaders was evaluated. We investigated whether comments were made with regard to experiences of children with SERDA. In addition, we evaluated and tried to remedy issues that emerged.

# Results

## Descriptives

SERDA was administered to 653 children, resulting in 176.6 hours of speech data. In addition, 569 and 622 classifications were obtained for the DMT and AVI respectively.

The usability of SERDA's tasks was evaluated by comparing their administration duration to the administration duration of the DMT and AVI. The administration of SERDA's word reading task took about 10 minutes on average, while the average duration of the passage reading tasks was 6 minutes. Administering both reading tasks generally took between 10 to 25 minutes, with an average duration of about 16 minutes. For the DMT and AVI, the COTAN reported that the combined duration of administration, scoring and interpretation are, respectively, 20 and 25 minutes. Thus, the total duration of administering SERDA's word and passage reading tasks is feasibly lower compared to the DMT and AVI, especially when both types of reading need to be administered, scored and interpreted.

During the word reading task children, on average, read 31 (SD = 8.6) words correctly, 24 (SD = 7.2) words per minute and 16 (SD = 7.4) words correctly per minute. For the passage reading task, children averaged 84 (SD = 19.6) words read correctly, 98 (SD = 34.7) words per minute, and 48 (SD = 20.1) words read correctly per minute. Figure 4 presents the distribution of administration duration, as well as the average accuracy, speed and WCPM measures for the word reading task. Likewise, Figure 5 presents the distribution of administration duration, as well as the average accuracy, speed and WCPM measures for the passage reading task.

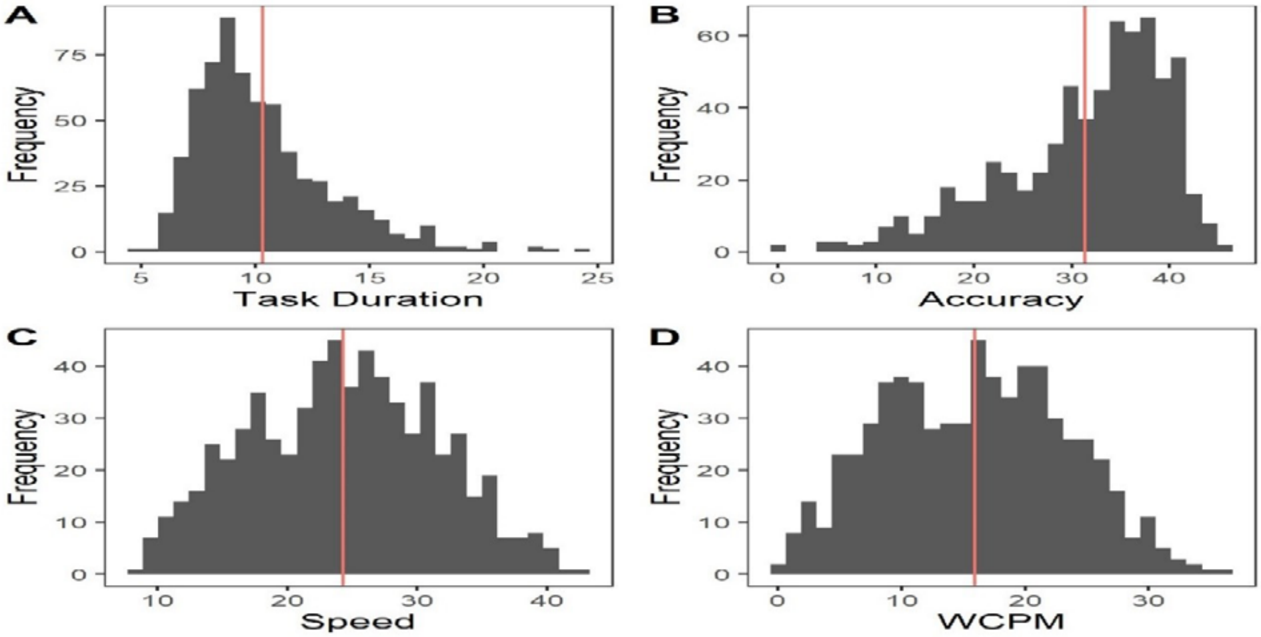


Figure 4 Average Task Duration (A), Accuracy (B), Speed (C) and WCPM (D) Metrics for the Word-reading Task, With Sample Averages Indicated by the Vertical Lines

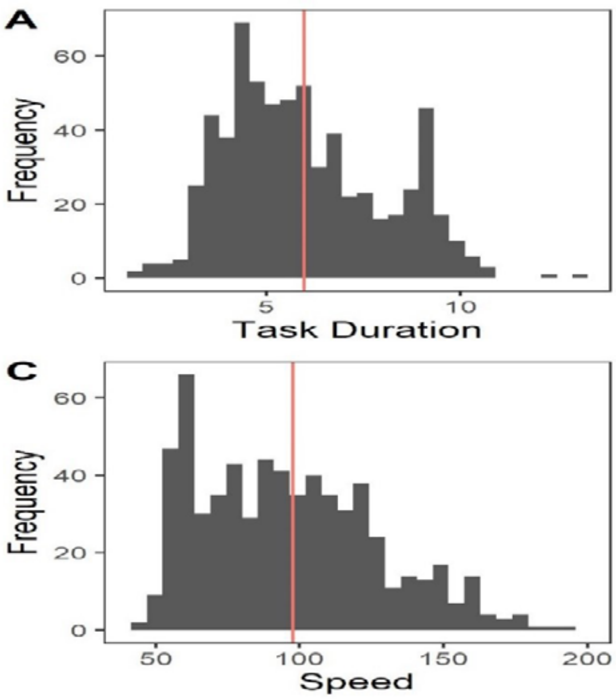


Figure 5 Average Task Duration (A), Accuracy (B), Speed (C) and WCPM (D) Metrics for the Passage-reading Task, With Sample Averages Indicated by the Vertical Lines

## Internal consistency

Table 2 shows Cronbach's Alpha for the accuracy scores of the word- and passage reading tasks, as well as their subtasks. Cronbach Alpha ranged from 0.89 to 0.97 for the subtasks, and between 0.96 to 0.98 for the complete tasks. This indicates that SERDA's reading tasks, when administered in their entirety, provide good internal consistency when used to make important individual decisions according to the COTAN guidelines (Evers et al., 2009).

| Task | Word-reading | Passage-reading |
|---|---|---|
| Complete | 0.96 [0.96, 0.97] | 0.98 [0.98, 0.99] |
| Sub-task 1 | 0.94 [0.93, 0.95] | 0.97 [0.96, 0.97] |
| Sub-task 2 | 0.89 [0.87, 0.91] | 0.96 [0.96, 0.97] |
| Sub-task 3 | 0.94 [0.93, 0.95] | 0.96 [0.96, 0.97] |

Note. Bracketed numbers are 95% confidence intervals.

Table 2 Cronbach's Alpha for the Accuracy Scores of the Word- and Passage Reading Tasks, as Well as Their Sub-tasks

## Split-half reliability

Table 3 shows the split-half reliability estimates of SERDA's word- and passage reading tasks, averaged over 10.000 random 50/50 splits. Average split-half reliability estimates ranged from 0.92 to 0.99, showing good reliability for the word- and passage reading task, when used for important individual decisions.

| Task-type | Words | Passages |
|---|---|---|
| Speed | 0.99 (SD < 0.01) | 0.93 (SD = 0.01) |
| Accuracy | 0.93 (SD = 0.01) | 0.97 (SD = 0.01) |
| WCPM | 0.97 (SD < 0.01) | 0.92 (SD = 0.01) |

Note. SD = standard deviation over 10000 split-half reliability estimates.

Table 3 Split-half Reliability for the Speed, Accuracy, and WCPM Measures of the Complete Word- and Passage Reading Tasks, Averaged Over 10000 Randomly Selected Splits

## Construct validity

To evaluate the construct validity of SERDA's reading tasks, we conducted a Pearson correlation between the WCPM-scores of the word- and passage reading tasks. In addition, we calculated spearman's rho between the WCPM-scores of the word reading task and the DMT classifications, as well as between the WCPM-scores of the passage reading task and the AVI classifications.

As shown in table 4, correlations varied between 0.54 to 0.81, showing moderate to strong positive relationships (Schober et al., 2018). The Pearson correlation analysis showed a significant positive relationship between the WCPM scores of SERDA's word and passage reading task, r(644) = .79, p < .001. The spearman correlation between the WCPM scores of the word reading task and the classification of the DMT showed a significant positive relationship, rho(502) = 0.54, p < .001. The spearman correlation between the WCPM scores of the passage reading task and the classification of the AVI showed a significant positive relationship, rho(591) = 0.81, p < .001.

| Task | WCPM Words | WCPM Passages | CLASS DMT | CLASS AVI |
|---|---|---|---|---|
| WCPM Words | 1 | - | - | - |
| WCPM Passages | 0.79 | 1 | - | - |
| Class DMT | 0.54 | 0.68 | 1 | - |
| Class AVI | 0.68 | 0.81 | 0.65 | 1 |

Note. All correlations were significant at a = 0.001

Table 4 Correlations Between the WCPM-scores of SERDA's Word- and Passage Reading Tasks, and the Performance Categories of the DMT and AVI

## Experiences with SERDA

Test leaders noted that children tended to enjoy SERDA's reading tasks, as exemplified through comments such as "I wasn't sure whether I should let her do the third set of words, because she hardly said anything right, but she enjoyed doing it so much that I thought it was fine", and " [I] experienced the task as very enjoyable". At the same time, some deemed the number of tasks to numerous: "She thought 3 word lists was a lot.", while others got tired: "had too little energy + concentration to finish the last story".

Regarding testing-issues, we observed some technical, practical and task-related problems. The test-leaders noted that children incidentally skipped tasks, while the application failed to store the recordings. Though these issues were mostly resolved by the end of the first round of data collection, this has led to some data loss. In addition, comments were made regarding disturbances on-site. For example, one test leader remarked: "There was a parent arguing in the hallway, which was quite distracting", while another noted that "the class next door was singing". Though these disturbances are problematic, as they reduce the quality of the audio recordings, they are deemed characteristic of primary schools. Finally, two issues were noted regarding children's performance on the word reading task. Namely, children "clicked too early", leaving them unable to recognise and read out the word, or read words before tapping the screen, leading to larger flashing times.

# Discussion

The current study aimed to improve reading fluency assessment by developing a novel digital reading fluency assessment instrument that utilizes ASR. Specifically, the goal of SERDA is to reduce the testing burden placed on teachers, while increasing the amount of available fluency diagnostics. Throughout this chapter we have investigated the usability, reliability and validity of SERDA's speed, accuracy and automaticity measures.

The results of the current study illustrate some advantages of SERDA compared to the DMT and AVI, based on its mode of administration. First of all, SERDA's administration time is relatively short. On average, administering both reading tasks takes about 16 minutes, whereas the combined administration time for the DMT and AVI comes down to around 20 minutes (COTAN; Egberink & Leng, 2024a, 2024b). In addition, SERDA only requires teachers to provide instructions and a microphone, while the DMT and AVI require test administration, scoring and interpretation to be done by hand. In practice, this can lead to total administration durations of up to 45 minutes. All the while, SERDA's usage of speech data allows for a more elaborate investigation of children's strengths and difficulties, as speed and accuracy information can easily and quickly be obtained at the item, task, and person level. Indeed, as long as tasks are conducted correctly and the speech of the child is properly captured, SERDA can be administered more quickly, reducing teacher's testing burden while providing detailed information on the speed and accuracy of children's reading.

Furthermore, the results of the current study indicate that both the word- and passage reading task provide reliable scores that resemble their pen-and-paper contemporaries moderately well to good. It is important to note, however, that the validity of the word-reading task was lower than the validity of the passage-reading task. This is not surprising, given that SERDA's word-reading task used a progressive demasking design. This design creates an administrative discrepancy with the DMT that reaches beyond the difference between the passage-reading task and the AVI, which primarily reflect their administrative modes. At the same time, the correlation between SERDA's word-reading task and the AVI was almost identical to the correlation between the DMT and AVI, indicating that SERDA's word-reading scores do not resemble the AVI's conceptually worse than the DMT's.

Though these results are promising, SERDA still requires work before it can solve the issues that currently plague fluency assessment. First of all, future research should be conducted to transform SERDA's provision of individual performance information into relevant diagnostics, as well as their translation into feedback towards teachers. An important step towards this goal can be made by incorporating the types of mistakes that children make at the item, task and person level. These mistakes and successes could then be used to discover reading profiles, based on state of the art learning-to-read models such as the DRC (Coltheart et al., 2001), Triangle (Harm & Seidenberg, 2004) and Connectionist Dual Process model (Perry et al., 2010). In turn, these reading profiles could be used to provide teachers with insightful individualized suggestions with regard to children's learning-to-read trajectories, an approach that has successfully been applied in the Netherlands within the context of reading comprehension (Keuning et al., 2019).

When more elaborate fluency diagnostics are incorporated, it is advised to undertake a more thorough validation of SERDA's reading tasks. Though the present study provided some evidence for the usability, reliability and validity of SERDA's reading tasks, more evidence is required if statements are to be made about the use of SERDA's speed, accuracy and WCPM scores in practice. Within this validation, extensive efforts should be made to evaluate the validity of the ASR-algorithm used throughout the current study. In addition, a more thorough validation could provide insights with regard to the reliability of score-estimation for children of different reading abilities, as well as a more elaborate evaluation of the quality and informativeness of individual items.

When a more elaborate validation of SERDA's improved measures has been established, researchers are advised to focus on the generation, evaluation and validation of prosody measures. Although some work has been conducted on extracting prosody measures using ASR (e,g. Truong et al., 2018), little research has been done to evaluate their usefulness and informativeness within a Dutch context, let alone for Dutch primary school children. Therefore, an elaborate investigation of the usefulness and validity of extracting established prosody measures from Dutch primary school children's speech should be conducted.

Finally, the comments made by the test-leaders require consideration. While children enjoyed the reading tasks, some found them too lengthy, reducing their ability to concentrate. However, the complete set of tasks was primarily administered for the purpose of the task evaluation. For applications of SERDA in practice, a subset of tasks could suffice, as substantiated by the split-half reliability results. In addition, comments were made regarding the tapping behavior of children, leading to errors at the item-level. These errors could be attributed to the novelty of the task. Therefore, we suggest a more thorough practice exercise to reduce these procedural errors.

To sum up, the current study was conducted with the goal of creating a reading fluency assessment tool that overcomes current assessment shortcoming. To achieve this goal, digital word- and passage reading tasks were developed based on expert opinion and currently popular reading fluency assessment instruments in the Netherlands. The results of this study suggest that SERDA's reading tasks provide reliable and valid indications of children's reading speed and accuracy, while reducing teacher's testing burden. Future researchers are advised to build upon the work conducted here by increasing the diagnostics SERDA can provide, through conducting a more comprehensive validation study on SERDA's reading tasks, and through the extraction of prosody features from the available speech data. If SERDA's development is successfully completed, its tasks could help individualize reading instruction, further reducing teacher testing burden, and improve reading education, providing the means to reduce the emerging emaciation of children's literacy the world over.

## References

Aldhanhani, Z. R., & Abu-Ayyash, E. A. (2020). Theories and Research on Oral Reading Fluency: What Is Needed?. *Theory and Practice in Language Studies*, *10*(4), 379–388. http://dx.doi.org/10.17507/tpls.1004.05

American Educational Research Association, American Psychological Association, & the National Council on Measurement in Education. (2014). *Standards for Educational & Psychological Testing*. Washington, DC: American Educational Research Association.https://www.testingstandards.net/uploads/7/6/6/4/76643089/standards_2014edition.pdf

Bai, Y., Hubers, F., Cucchiarini, C., & Strik, H. (2020). *ASR-Based Evaluation and Feedback for Individualized Reading Practice*. Paper presented at Interspeech 2020, Shanghai, China, October 25–29 [online]. Available: https://www.researchgate.net [March 21, 2024].

Bai, Y., Hubers, F. C. W., Cucchiarini, C., & Strik, H. (2021). *An ASR-based Reading Tutor for Practicing Reading Skills in the First Grade: Improving Performance through Threshold Adjustment*. Paper presented at Iberspeech 2021, Valladolid, Spain, March 24-25 [online]. Available: https://repository.ubn.ru.nl/bitstream/handle/2066/245151/245151.pdf [March 21, 2024]

Benjamin, R. G., Schwanenflugel, P. J., Meisinger, E. B., Groff, C., Kuhn, M. R., & Steiner, L. (2013). A spectrographically grounded scale for evaluating reading expressiveness. *Reading Research Quarterly*, *48*(2), 105–133. https://doi.org/10.1002/rrq.43

Bolaños, D., Cole, R. A., Ward, W. H., Tindal, G. A., Hasbrouck, J., & Schwanenflugel, P. J. (2013). Human and automated assessment of oral reading fluency. *Journal of Educational Psychology, 105*(4), 1142–1151. https://doi.org/10.1037/a0031479 \

Cheng, J. (2011). *Automatic assessment of prosody in high-stakes English tests*. Paper presented at Interspeech 2011, Florence, Italy, August 27-31 [online]. Available: https://caimber-cdn.s3.us-west-2.amazonaws.com/papers-and-presentations/cheng11c_interspeech.pdf [Januari 21, 2024].

Chung, W. L., & Bidelman, G. M. (2022). Acoustic Features of Oral Reading Prosody and the Relation With Reading Fluency and Reading Comprehension in Taiwanese Children. *Journal of Speech, Language, and Hearing Research*, *65*(1), 334–343. doi:10.1044/2021_JSLHR-21-00252

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). The DRC model: A model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–258.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Cucchiarini, C., van Hamme, H., Driesen, J., Sanders, E. (2008). *THE JASMIN-CGN: CORPUS Design, recording, transcription and structure of the corpus.*

Cucchiarini, C., Strik, H., & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *The Journal of the Acoustical Society of America*, *107*(2), 989–999. doi:10.1121/1.428279

Dimzon, F. D., & Pascual, R. M. (2023). Prosodic characterisation of children's Filipino read speech for oral reading fluency assessment. *International Journal of Technology Enhanced Learning*, *15*(1), 74–94. doi:10.1504/ijtel.2023.127939

Egberink, I.J.L. & Leng, W.E. de. (2024a). *2010, AVI [2010, AVI]* [online]. Available: www.cotandocumentatie.nl [Januari 21, 2024].

Egberink, I.J.L. & Leng, W.E. de. (2024b). *2010, Drie-Minuten-Toets* [2010, Three Minute Test] [online]. Available: www.cotandocumentatie.nl [Januari 21, 2024].

Elffers, B., van Bael, C., Strik, H. (2005). *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions* [online]. Available: https://hstrik.ruhosting.nl/wordpress/wp-content/uploads/2013/03/a121-adapt.pdf

Evers, A., Lucassen, W., Meijer, R., & Sijtsma, K. (2009). *COTAN beoordelingssysteem voor de kwaliteit van tests (geheel herziene versie)* [COTAN evaluation system for the quality of tests (completely revised version)]. NIP. Available: http://www.psynip.nl/website/wat-doet-het nip/tests/beoordelingsprocedure/beoordelingsprocedure [Januari 21, 2024].

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: the dtw package. *Journal of statistical Software*, *31*, 1-24. Doi: 10.18637/jss.v031.i07

Godde, E., Bailly, G., Escudero, D., Bosse, M. L., & Gillet-Perret, E. (2017). *Evaluation of reading performance of primary school children: Objective measurements vs. subjective ratings*. Presented at the 6th Workshop on Child Computer Interaction, Glasgow, Scotland, November 13 [online]. Available: doi:10.21437/WOCCI.2017-4

Grainger, J., & Segui, J. (1990). Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Perception & psychophysics*, *47*, 191-198. https://doi.org/10.3758/BF03205983

Gubbels, J., van Langen, A., Maassen, N., & Meelissen, M. (2019). *Resultaten PISA-2018 in vogelvlucht* [Results of PISA-2018 from a bird's eye view] [online]. Enschede: Universiteit Twente. Available: https://doi.org/10.3990/1.9789036549226 [Januari 21, 2024].

Harm, M. W., & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*, 662–720. doi:10.1037/0033- 295X.111.3.662

Horning, A. S. (2007). Reading across the curriculum as the key to student success. *Across the disciplines*, *4*(1), 1–17.

Inspectie van het Onderwijs. (2024). *Schoolweging primair onderwijs* [Schoolweightprimary education] [online]. Available: https://www.onderwijsinspectie.nl/trends-en-ontwikkelingen/onderwijsdata/schoolweging-po [March 20, 2024].

Keuning, J., Swart, N., Scheltinga, F., Gruhn, C.S., Segers, E. & Verhoeven, L. (2019). *Evaluatie en planning van leesleertrajecten: Een dynamisch perspectief (eindrapport NRO project 405-15-548)* [Evaluation and planning of learning-to-read trajectories, a dynamic perspective (final report NRO project 405-15-548)] [online]. Arnhem: Cito. Available: https://www.nro.nl/sites/nro/files/migrate/eindrapport-405-15-548.pdf [Januari 21, 2024].

Keuning, J. & Verhoeven, L. (2005). *Signaleren van lees- en spellingproblemen in groep 4-8* [Detection of reading- and spelling problems in Grades 2-6]. Nijmegen: Expertisecentrum Nederlands

Kim, Y. S. G., Quinn, J. M., & Petscher, Y. (2021). What is text reading fluency and is it a predictor or an outcome of reading comprehension? A longitudinal investigation. *Developmental Psychology*, *57*(5), 718–732. https://doi.org/10.1037%2Fdev0001167

Kuhn, M., Schwanenflugel, P. & Meisinger, E. (2010). Aligning theory and assessment of reading fluency: Automaticity, prosody, and definitions of fluency. *Reading Research Quarterly, 45*(2), 232–253. https://doi.org/10.1598/RRQ.45.2.4

Levis, J., & Suvorov, R. (2012). Automatic speech recognition. In: Chapelle, C. A. (2012). *The encyclopedia of applied linguistics.* Hoboken, NJ : John Wiley & Sons.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Loukina, A., Klebanov, B. B., Lange, P. L., Qian, Y., Gyawali, B., Madnani, N., Misra, A., Zechner, K., Wang, Z., & Sabatini, J. (2019). *Automated Estimation of Oral Reading Fluency During Summer Camp e-Book Reading with MyTurnToRead.* Paper presented at Interspeech 2019, Graz, Austria, 15-19 September [online]. Available: https://www.iscaarchive.org/interspeech_2019/loukina19_interspeech.pdf [Januari 21, 2024].

Louradour, J. (2023). *Whisper-timestamped* [online]. GitHub Repository. Available:https://github.com/linto-ai/whisper-timestamped [March 21, 2024].

Meelissen, M. R. M., Maassen, N. A. M., Gubbels, J., van Langen, A. M. L., Valk, J., Dood, C., Derks, I., In 't Zandt, M., & Wolbers, M. (2023). *Resultaten PISA-2022 in vogelvlucht* [Results of PISA-2022 from a bird's eye view] [online]. Enschede: Universiteit Twente. Available: [March 18, 2024].

Morris, D., & Perney, J. (2018). Using a sight word measure to predict reading fluency problems in grades 1 to 3. *Reading & Writing Quarterly*, *34*(4), 338–348. https://doi.org/10.1080/10573569.2018.1446857

Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., & Tobin, B. (2003). Evaluation of an automated reading tutor that listens: Comparison to human tutoring and classroom instruction. *Journal of Educational Computing Research*, *29*, 61–117. https://doi.org/10.2190/06AX-QW99-EQ5G-RDCF

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading* [online]. Boston College, TIMSS & PIRLS International Study Center. Available: https://doi.org/10.6017/lse.tpisc.tr2103.kb5342 [Januari 21, 2024].

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: U.S. Government Printing Office.

Nicolao, M., Sanders, M., & Hain, T. (2018). *Improved acoustic modelling for automatic literacy assessment of children*. Paper presented at Interspeech 2018, Hyderabad, India, 2-6 September [online]. Available: https://doi.org/10.21437/Interspeech.2018-2118 [Januari 21, 2024].

Paige, D. D., Rupley, W. H., Smith, G. S., Rasinski, T. V., Nichols, W., & Magpuri-Lavell, T. (2017). Is prosodic reading a strategy for comprehension?. *Journal for educational research online*, *9*(2), 245–275. Doi:10.25656/01:14951

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*, 106–151. https://doi.org/10.1016/j.cogpsych.2010.04.001

Pikulski, J.J., & Chard, D.J. (2005). Fluency: Bridge between decoding comprehension. *The Reading Teacher, 58*(6), 510–519. https://doi.org/10.1598/RT.58.6.2

Posit team (2023). RStudio: Integrated Development Environment for R, version 4.3.1. Posit

Software, PBC, Boston, MA. http://www.posit.co/.

Proença, J., Celorico, D., Candeias, S., Lopes, C., & Perdigão, F. (2015). *Children's Reading Aloud Performance: A Database and Automatic Detection of Disfluencies.* Presented at Interspeech 2015, Dresden, Germany, 6-10 September [online].

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). *Robust speech recognition via large-scale weak supervision.* Paper presented at the 40th International Conference on Machine Learning, Hawaii, USA, 23-29 July [online]. Available: https://proceedings.mlr.press/v202/radford23a.html [March 21, 2024].

Reeder, K., Shapiro, J., & Wakefield, J. (2007). *The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for Canadian immigrant children*. Proceedings of the 9th European Conference on Reading [online].

Reynolds, C. R., Livingston, R. A., & Allen, D. N. (2021). *Mastering modern psychological testing: Theory & methods (Second Edition).* Cham: Springer Nature Switzerland.

Sabu, K., & Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Transactions on ICT*, *6*, 221–225. https://doi.org/10.1007/s40012-018-0202-3

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763–1768. DOI: 10.1213/ANE.0000000000002864

Silva, W. A., Carchedi, L. C., Junior, J. G., de Souza, J. V., Barrere, E., & de Souza, J. F. (2021). A framework for large-scale automatic fluency assessment. *International Journal of Distance Education Technologies (IJDET)*, *19*(3), 70–88. DOI: 10.4018/IJDET.2021070105

Society for Learning Analytics. (2011). *What is Learning Analytics?* [online]. Available:https://www.solaresearch.org/about/what-is-learning-analytics/ [Januari 21, 2024].

Swart, N. M., Gubbels, J., in 't Zandt, M., Wolbers, M. H. J., & Segers, E. (2023). *PIRLS-2021: Trends in leesprestaties, leesattitude en leesgedrag van tienjarigen uit Nederland* [online] [PIRLS-2021: Trends in the reading performance, reading attitude and reading behaviour of ten year olds from the Netherlands]. Expertisecentrum Nederlands. Available: PIRLS-2021_Rapportage.pdf(expertisecentrumnederlands.nl) [Januari 21, 2024].

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., Vloedgraven, J. & De Wijs, A. (2018a). *Wetenschappelijke verantwoording DMT* [Scientific justification DMT] [online]. Cito: Arnhem. Available: https://cito.nl/media/2juinmdl/106-cito_lvs-dmt_gr-3-tm-halverwege-gr-8_wet-verantwoording.pdf [Januari 21, 2024].

Van Til, A., Kamphuis, F., Keuning, J., Gijsel, M., & De Wijs, A. (2018b). *Wetenschappelijke verantwoording AVI* [Scientific justification AVI] [online]. Cito: Arnhem. Available: https://cito.nl/media/hdqkk1if/109-cito_lvs-avi_gr-3-tm-halverwege-gr-8-wet-verantwoording.pdf [Januari 21, 2024].

Truong, Q. T., Kato, T., & Yamamoto, S. (2018). *Automatic Assessment of L2 English Word Prosody Using Weighted Distances of F0 and Intensity Contours.* Paper presented atInterspeech 2018, Hyderabad, India, 2-6 September [online].

University of Oregon (2020). *8th Edition of Dynamic Indicators of Basic Early Literacy Skills (DIBELS®): Administration and Scoring Guide*. Eugene, OR: University of Oregon.

https://dibels.uoregon.edu

Veenendaal, N.J., Groen, M.A., & Verhoeven, L. (2016). Bidirectional Relations between Text Reading Prosody and Reading Comprehension in the Upper Primary School Grades: A Longitudinal Perspective. *Scientific Studies of Reading. 20*(3), 189–202.

https://doi.org/10.1080/10888438.2015.1128939

Van der Velde, M. E., Molenaar, B., Veldkamp, B. P., Feskens, R. C. W., Keuning, J. (2024). *What do They say? Assessment of Oral Reading Fluency in Early Primary School Children: A Scoping Review*. Manuscript submitted for publication.

Wei, X., Cucchiarini, C., van Hout, R. W. N. M., & Strik, H. (2022). Automatic Speech Recognition and Pronunciation Error Detection of Dutch Non-native Speech: cumulating speech resources in a pluricentric language. *Speech Communication*, *144*, 1–9. https://doi.org/10.1016/j.specom.2022.08.004

Yilmaz, E., & Pelemans, J., van Hamme, H. (2014). *Automatic assessment of children's reading with the FLaVoR decoding using a phone confusion model.* Paper presented at Interspeech 2014, Singapore, 14-18 September [online].

# Norway

# NORWAY: BIOGRAPHIES

## Kevin Steinman

After working as an assistant professor in English literature, area studies and didactics (University of Oslo, Inland Norway University of Applied Sciences), Kevin now coordinates Norway's summative assessments in English and English for students with sign language. As senior adviser at Norway's Directorate for Education and Training, he has helped to lead Norway's digitization of centrally administered English examinations since 2020.

Kevin is the co-author of a textbook for lower secondary English, and wrote a chapter in Poetry and Sustainability in Education (Kleppe and Sorby, eds., 2022), on using Indigenous poetry from North America for teaching sustainability in upper secondary English.

Before moving to Norway, Kevin worked in the US as a full-time musical performer, producer and recording artist.

# NORWAY'S 2024 GENERATIVE AI JOURNEY IN SCHOOLS AND ASSESSMENTS: RESPONDING TO A CALL FOR GUIDANCE ON NEWER DIGITAL TECHNOLOGIES IN NORWAY'S SECONDARY SCHOOLS

## Abstract

This article presents Norway's exploration of both the challenges and opportunities teachers and pupils are experiencing since Large Language Models' (LLMs') sudden entry into the classroom two years ago. By establishing new guidance, the Directorate's role as advising authority is highlighted in the article's first section, along with the challenges of advising the education sector in the early stages of ongoing technological revolutions. The second section details the Directorate's response to generative AI's prospective impact on centralized, written assessments. The article concludes by unpacking the recent digitalization of the high-stakes, secondary English written assessment.

## Background

With the increased availability of GenAI tools like ChatGPT, an LLM made publicly available in November, 2022, pupils, teachers and school administrators are in the process of integrating new digital tools into learning practices. While evidence informed practice (EIP) has long been touted as an "essential feature of effective education systems" (Mincu 2014; Greany 2015, in Nelson & Campell 2017), evidence for clear guidance or best practice when it comes to GenAI in the classroom is still in the early stages of being collected and analysed (Fullan, et al. 2023).[1] Like during the COVID-19 pandemic, where governments and schools felt their way forward in an extreme case of (digitized) learning by doing, LLMs and other GenAI tools present the education sector with a myriad of new and pressing challenges to solve – in media res.

The challenges presented by GenAI come with intriguing opportunities. Early recommendations call for "effective, ethical and collaborative" solutions (Russel Group, 2023), given that GenAI is "widely available, ... likely only to become more sophisticated, and has both specific negative and

[1] The Directorate for Education and Training's webpage on Advice for use of Artificial Intelligence in schools acknowledges this clearly: "praxis, knowledge development and laws will lag behind [the development of technology]" (author's translation). https://www.udir.no/kvalitet-og-kompetanse/digitalisering/kunstig-intelligens-ki-i-skolen/

unique positive potential for education” (Miao & Holmes, 2024, emphasis added). Particularly compelling is the duality of this call. Pedagogical challenges range from how to identify authentic digital student-authored texts – if, indeed, any such artifact still exists -, to the potential threats to democratic ideals posed by the biases of LLMs, which can often be presented as fact. It also presents challenges to the administration of summative assessments, as well as data privacy. Finally, while this paradigm shift brings risks of widening the digital divide, it also brings with it power-shifting opportunities reminiscent of the invention of the printing press.

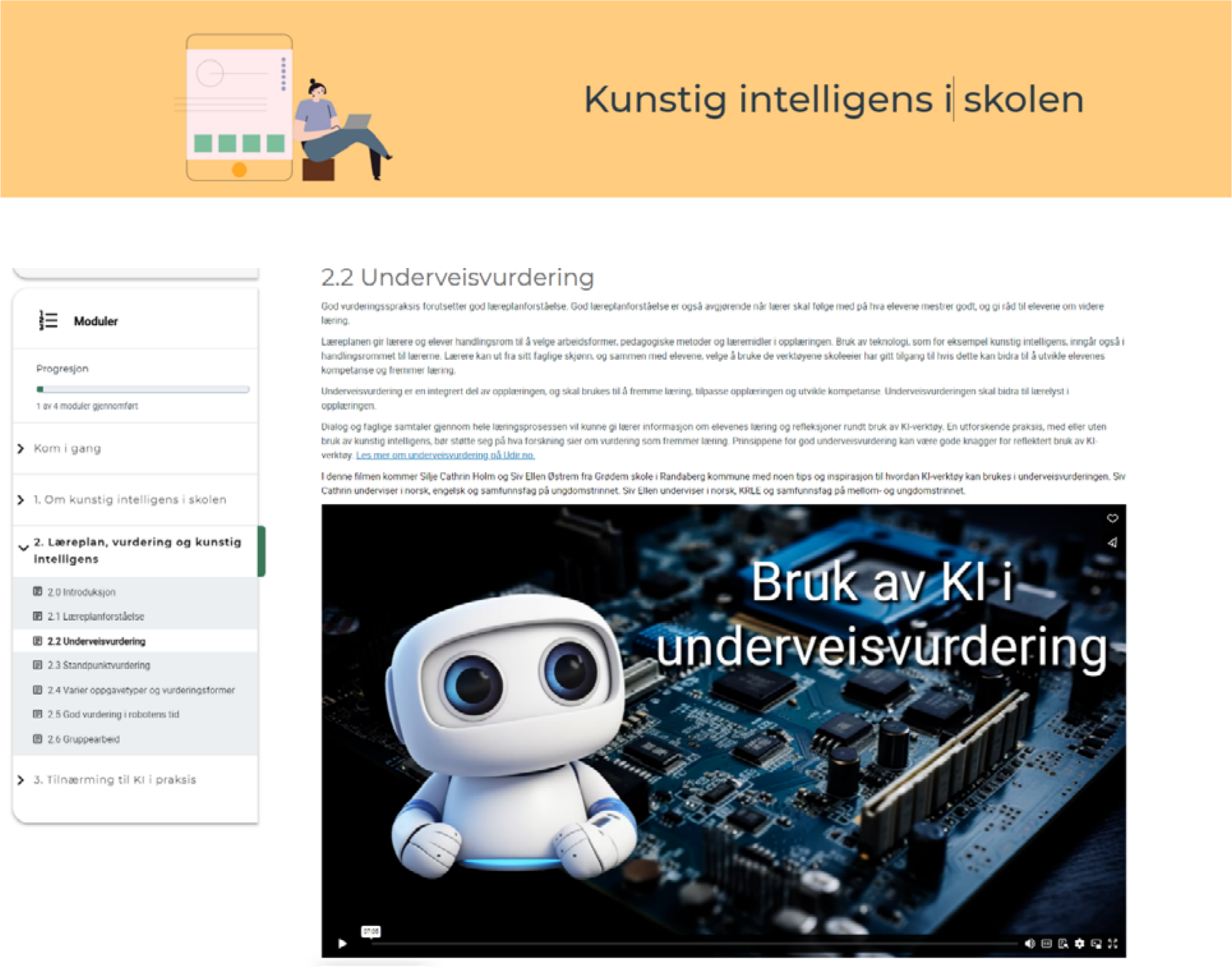
The introduction of widely accessible generative AI has been described as a factor that may “radically reshape and redefine the nature of learning and teaching” (Dobrin 2023 in Fullan, et al, 2024). As of April 2024, more than half of Oslo schools report that they are using GenAI in their instruction (Røyne, 2024, Gerhardsen, 2024). Randaberg, near Stavanger, launched its own version of ChatGPT for use in schools in 2023. In many other places in Norway, uptake and practices vary.

Norway’s Directorate of Education and Training has responded to the call for guidance on the question of how to use LLMs pedagogically in several ways. This guidance can be seen as a resource for teachers and school administrators developing their professional digital competence (PDC). PDC has been a focus in recent years (Nagel, 2024; Udir, 2018), a situation which has only increased in importance in the wake of the Covid-19 pandemic (Garcia, et al. 2022).

## Support in the classroom

### Competence package

In February 2023 Norway’s Directorate of Education and Training published a digital resource (kompetansepakke - “competence package”) on using AI in schools. Hosted on the Directorate’s website, it requires login through one of two national authentication systems. Its modules cover teaching practice, assessment, source criticism, protection of individuals’ data, among other things; and its content was designed to help school administrators and teachers navigate this technology’s integration into the classroom. Its form follows a popular mode of pedagogical professional development support provided by the Directorate: a blend of multimedia, written and graphic information, building up to discussion tasks as well as other activities. As of April 2024, over 6,000 individual users had made use of the resource, demonstrating a relatively high level of interest in this topic.


Kunstig intelligens i skolen
Moduler
Progresjon
1 av 4 moduler gjennomført
Kom i gang
1. Om kunstig intelligens i skolen
2. Læreplan, vurdering og kunstig intelligens
2.0 Introduksjon
2.1 Læreplanforståelse
2.2 Underveisvurdering
2.3 Standpunktvurdering
2.4 Varier oppgavetyper og vurderingsformer
2.5 God vurdering i robotens tid
2.6 Gruppearbeid
3. Tilnærming til KI i praksis
Bruk av KI i underveisvurdering

2.2 Underveisvurdering

God vurderingsspraksis forutsetter god læreplanforståelse. God læreplanforståelse er også avgjørende når lærer skal følge med på hva elevene mestrer godt, og gi råd til elevene om videre læring.

Læreplanen gir lærere og elever handlingsrom til å velge arbeidsformer, pedagogiske metoder og læremidler i opplæringen. Bruk av teknologi, som for eksempel kunstig intelligens, inngår også i handlingsrommet til lærerne. Lærere kan ut fra sitt faglige skjønn, og sammen med elevene, velge å bruke de verktøyene skoleeier har gitt tilgang til hvis dette kan bidra til å utvikle elevenes kompetanse og fremmer læring.

Underveisvurdering er en integrert del av opplæringen, og skal brukes til å fremme læring, tilpasse opplæringen og utvikle kompetanse. Underveisvurderingen skal bidra til lærelyst i opplæringen.

Dialog og faglige samtaler gjennom hele læringsprosessen vil kunne gi lærer informasjon om elevenes læring og refleksjoner rundt bruk av KI-verktøy. En utforskende praksis, med eller uten bruk av kunstig intelligens, bør støtte seg på hva forskning sier om vurdering som fremmer læring. Prinsippene for god underveisvurdering kan være gode knagger for reflektert bruk av KI-verktøy. Les mer om underveisvurdering på Udir.no.

I denne filmen kommer Silje Cathrin Holm og Siv Ellen Østrem fra Grødem skole i Randaberg kommune med noen tips og inspirasjon til hvordan KI-verktøy kan brukes i underveisvurderingen. Siv Cathrin underviser i norsk, engelsk og samfunnsfag på ungdomstrinnet. Siv Ellen underviser i norsk, KRLE og samfunnsfag på mellom- og ungdomstrinnet.

Figure 1: Artificial Intelligence in schools: 2.2 Assessment for learning

(Utdanningsdirektoratet, 2024)

The competence package encourages users to spend time in professional dialogue unpacking the nature and use cases of GenAI, giving examples of constructive pedagogical uses of tools like ChatGPT – as well as less optimal ones. This background work prepares teachers and administrators to go into more reflective dialogue on the questions surrounding challenges in helping pupils familiarize themselves with these benefits and pitfalls.

This AI in Schools competence package was updated in January 2024, with an expanded section focusing on assessment. The new module focused on collecting evidence of learning, especially in assessment for learning situations, as well as sharpening curricular understanding, and the use of AI in assessment situations. In July the Directorate released a thoroughly revised updated version of the competence package, filling out and strengthening the sections on assessment, among other things. This revision includes whole new sections, such as subject-specific advice for the use of AI in the classroom, with practical examples. In addition, a new landing page on the Directorate's website was launched (Kunstig intelligens i skolen | udir.no), helping the education sector navigate this hot topic issue. This is very much an ongoing process, with updates expected as learning-centered use of AI becomes more and more widespread.

## Roundtable on AI in education

In February 2024, a roundtable conference was hosted in Oslo by the Directorate of Education and Training, with the goal of establishing dialogue between researchers, policy makers, teacher and school administration organizations and Ministry of Education officials on the questions surrounding AI in schools. This event can be seen as "stakeholder consultation", a support for higher quality use of evidence in education systems (Rickinson, et al, 2022). This forum offered participants an opportunity to, among other things, give feedback to the Directorate on its AI in Schools competence package, as well as learn about several ongoing research projects in their early stages, and prognosticate on the future of AI in schooling.

The research projects that attendees learned about ranged from a public/private project researching trustworthiness in the use of AI for education: EduTrust AI (Wasson, et al., 2024); to a project drawing on university and municipality resources to investigate "integrating the didactic use of artificial intelligence into the established structure with competence development for teachers in Asker" (Gilje, 2024). In addition Østfold University College presented their public/ private research into developing a new pedagogical tool for formative feedback in English process writing (Engeness, et al., 2024). Finally, a project in Vestfold County was announced, showing that research into pedagogical uses of AI is not only being generated by tertiary education faculties, but also by regional governments. When completed, these research projects will lay the groundwork for an even more solid knowledge base.

In discussion groups, where questions ranged from the theoretical to the practical, the Directorate's representatives – from several different departments – took on a secretarial role. Advisers from the Directorate were tasked with facilitating conversations between stakeholders, and taking notes of questions and concerns that arose. Notes taken by Directorate representatives would be fed back into planning for future communications/ support for the educational community, as well as future events like this roundtable.

One of the clearest messages we heard was that, while the competence package was appreciated for its timeliness, many teachers across the country desired more guidance on the best practice of AI in the classroom. In other words, while it is true that a timely launch of a competence package limited how extensive its evidence-based advice could be, it is also the case that classroom implementation was already happening before the competence package was launched. This feedback was very important for the Directorate's next steps.

## Advice on AI in schools

In February 2024 Norway's Directorate of Education and Training published new advice on the use of AI in schools ("*Råd om kunstig intelligens i skolen*"). While concerns about AI, whether environmental, ethical or pedagogical, are neither hard to identify nor understand, it is also clear that the technology's new possibilities are potentially transformative. Therefore the Directorate considers providing information about AI as part of the mandate of the education in Norway: to "open doors to the world and to the future".[2] This is the reason the directorate declares that schools must work with AI, even while guidelines are being more fully developed.

This advice-centered pedagogical support begins with an acknowledgement that practice and guidance on AI are lagging behind technology's advancements (Utdanningsdirektoratet, 2024); at the same time it makes clear that it is already possible to guidepost some important things to consider when integrating AI in the classroom.

The advice is framed by a mandate inviting involvement at all levels: "Technology is changing schools – together we must decide how" (Utdanningsdirektoratet, author's translation). In outlining its new AI advice, the Directorate highlights two questions:

1. "What kind of society and work life will be created by AI's swift changes?"
2. "What values will be important to emphasise in school, when technological advances contribute to ever more unstable source criticism?" (Udir, 2024, author's translations)

These two questions frame the Directorate's advice as important in addressing new and emerging challenges spanning collective Norwegian identity, its post-digital labour market and (digital) ethics.

[2] *„opne dører mot verda og framtida" (Schools Act § 1-1. Formålet med opplæringa [The purpose of education], author's translation).*

**Table 1. Advice on artificial intelligence in schools**

| |
|---|
| 1. **Point out relevant areas in the curriculum:** critical digital literacy, source criticism, ethics, personal data security and democracy. |
| 2. **The school and the teacher (must) assess when AI is relevant**, based on competence aims and content in the curriculum. The teacher is the pedagogical leader in the classroom, who stakes out good learning processes. |
| 3. **Emphasise variation in instruction and assessment.** Explore approaches that give the teacher assurances that it is the pupil's own competence that is being assessed. For example, assessment in various phases of a writing process. |
| 4. **Talk with pupils** about what AI is, what possibilities the technology allows, but also which dangers it brings. Explore good working processes and learning approaches together. |
| 5. **Talk with parents and guardians** about digital praxis and use of AI tools in the school and home. |
| 6. **Use professional learning communities** to tackle complex challenges which are difficult to tackle alone. School authorities and School leaders must prioritize development of professional digital competence. |
| 7. **Create a culture for trialing** and assessment of pedagogical practice. |
| 8. **Use secure solutions** assessed and approved by school authorities. |
| 9. **Consider pupils' age and maturity,** and demonstrate caution especially when it comes to younger children. Make use of the technological possibilities that exist to customize tools for pedagogical best practice. |

(Udir, 2024, author's translation)

While the Directorate recognizes that this advice can be deemed limited in scope, it is worth noting that some organisations have regretted early advice, given the speed with which GenAI has been rolled out and improved.

## Increasing support for summative assessment integrity

Summative assessment for secondary schools in Norway is both centrally and locally administered. The Directorate for Education and Training has responsibility for all centrally administered, written examinations. This involves especially their creation, construct validation and reporting. More than 300,000 pupils/ private scheme candidates take these centrally administered, written exams every year in hundreds of different subject codes.

These high-stakes, certification examinations, while made and delivered by central education authorities, are administered locally in partnership with regional authorities (Fylkeskommuner for upper secondary, Statsforvalter for lower secondary). Grading is done by teachers recruited from across the country in a process administered by County governors (Statsforvalter). In addition, the Directorate's exams are used in a separate strand of examinations for private scheme candidates, supporting non-traditional certification (without being enrolled in school), as well as those individuals who might wish to improve a previous grade. All these assessments require strong, effective communication and collaboration across all levels of the education sector.

Since 2020, the Directorate has digitized Norway's centrally administered upper secondary written exams. This means that the rise of GenAI comes on the heels of a three-year long process in which we have transformed a previously paper-behind-glass examinations system into a nearly fully digitized one.[3] While this process will be given more expansive treatment in the final section of this article, it is important to note here the relatively high costs in both financial resources and human efforts that were necessary to make this change happen.[4] Given these, any challenges that GenAI might present to the integrity of the new digital assessment system will most probably need to be solved in other ways than returning to paper-based, "semi-digital" solutions.

Of the 1000+ subject codes coordinated by Norway's Directorate of Education and Training, some 112 have so-called "preparation days", which are obligatory school days giving pupils the opportunity to study, reflect and prepare for their upcoming examination in a particular subject. It may be worth noting that, apart from year 13, when every student takes the Norwegian language arts and literature exam, as well as a few other advanced study subjects, pupils only find out which subject(s) they will be examined in two days prior to the exam(s). So the function of preparation day may well be seen in that context: as a helpful supplement to the instruction the student has received throughout the school year, as well as a useful opportunity to refresh, reflect and focus on what one has learned in the past year.

This policy of a requiring a full school day to work with textbooks, previous assignments, etc., presents graders with some interesting challenges when it comes to identifying texts written whole cloth before the exam day. Pupils are rightly encouraged to use some material they have

[3] There remain some subjects that deliver their exams via the old system, but for the most part, Norway has made this transition so that for more than 90% of the school subjects, fully digital exams now are the only option.

[4] It may also be worth noting that what could be called a large school-cultural change took place during digitization process, where traditions were challenged and updated. While this is not something that can in and of itself argue for maintaining digitization, it is a part of the process that should be recognized.

prepared before the exam – this is in fact central to the idea of preparation day. But as the purpose of examination is to discover the pupils' own competence,[5] it has never been thought of as an ideal outcome that school examinees submit entire products of previous writing sessions. Monitoring this problem, however, has traditionally proven tricky.

This inherited problem has become only more challenging in the wake of widely available generative AI, which at the stroke of a few keys can now generate as many texts on a subject as one might wish, stored digitally on the same computer with which pupils will take the exam. As this combination of factors poses the risk of potentially impacting assessment integrity negatively, the Directorate has informed the Department of Education that a public hearing is needed on the consequences of eliminating the preparation day. That hearing took place from Spring through early Fall 2024, collecting feedback from all interested stakeholders, with a decision slated in time for implementation, Spring 2025. The goal of the hearing is to find out what the education sector thinks about this situation, and to map the consequences of potentially eliminating the preparation day as an element of the assessment ecology.

In addition to this path of enquiry, the Directorate has made other recommendations for supporting the integrity of summative assessments, such as

- Improving regional routines for the proctoring of central written exams
- New routines for grading, especially when it comes to following up suspected malpractice
- Increasing pay for graders

The Directorate has at the same time undertaken a process by which the number of digital test-taking tools (digital dictionaries, digital encyclopedias, and the like) may be greatly reduced. In 2006 a digitization policy was enacted allowing pupils access to the same digital tools in assessment situations that they were familiar with in the classroom.

While this was seen as a forward-looking strategy at the time, and aligned with the 1-1 digital strategy for pupils, it was difficult for officials to have predicted that within 16 years a change as dramatic as GenAI would make such a student-centered, digital-focused approach so potentially challenging to manage.

Since these approved online assessment-taking tools are managed by regional authorities, any effort to trim them to a more manageable level must be coordinated carefully and transparently. This streamlining process to is already in motion, with the number having been pared down from 150 to 60. The number of digital tools available on exam day is thought to be possible to pare down ultimately to 20. While Norway's practices outlined above, collectively, might place Norway toward the "liberal" end of a scale measuring affordances supporting examination, it may be worth noting that that Norway scores high on international comparisons of the level of trust in its public officials. So it may be fair to say that a measure of this same trust has historically been extended to its pupils when it comes to summative assessment frameworks.

[5] «Eksamenskarakteren skal gi uttrykk for den individuelle kompetansen til kvar elev» («the examination grade shall give an impression of the individual competence of each pupil». Ministry of Education, Regulations to the Education Act §3-14, author's translation)

The education system expects pupils to demonstrate their progress in learning in a judicious and ethical way. How GenAI impacts this will be interesting to follow.

Already, some early results are intriguing. A February 2024 survey of 1,234 teachers found that 6 in 10 report having caught pupils passing off text produced by GenAI as their own (Molnes, 2024). This points to the highly dynamic nature of our present moment, in which norms and expectations are being thoroughly reviewed on an ongoing basis. Some teachers report returning to pen and paper when they need to evaluate student writing. Meanwhile, schools are adopting responsible ways of integrating GenAI into the classroom.

Oslo's school district (the country's largest), for instance, has launched its own version of ChatGPT, allowing pupils and teachers the chance to use it in classroom settings since the fall of 2023. Randaberg, near Stavanger, was one of the first schools in the country to launch its own version of ChatGPT, already in January, 2023. These are just two examples of many schools doing early exploratory work, encouraged by the advice mentioned above. While such explorations are clearly in line with the Directorate's goals of preparing pupils for success in the future, the many layers of administration between the Directorate and schools means that it is unfortunately difficult for officials who advise the education sector to maintain a clear picture of what is working, and what is not working. Still, the government as recently as June, 2024 has declared an intent to be more proactive.

## Norway's digitization of high-stakes upper secondary English exams

After detailing some of Norway's responses to newer challenges to ed-tech practices, it may be worth unpacking the digitization process secondary English summative assessments have undergone in recent years. This is especially relevant given the ways in which such assessments are being asked to take GenAI into account.

The redesign of the high-stakes English exam after curricular renewal in Norway (2020) has been described by observers and stakeholders as transformative. It is also one of the first high-stakes exams in Europe to be offered as a fully digital certification assessment. Traditionally an exclusively written assessment with two tasks (one short response, one long), the 10th and 11th grade[6] English exams have been expanded to include listening and reading comprehension, as well as three written response tasks (two short, one longer). Importantly, the exam now takes place in a new all-digital environment, where items are authored, quality assured, piloted, analysed, published, administered and graded on a digital platform maintained by the Directorate. This section discusses the changes involved in making this transition, ranging from new methods to improved curricular coverage, as well as some of the practical challenges and opportunities brought about by such changes.

[6] 11th grade English is offered in two parallel examinations: 1. vocational and 2. general studies

## Background

In 2020, an Expert Group delivered its recommendations to the Directorate for changes to the centrally administered, certification examinations in Norway. These included observations of interrater reliability below optimal levels in many subjects, especially language. Their findings called for improvements to the examination system, focusing on four central themes:

- Reliability
- Validity
- Sustainability
- Fairness

Importantly, the recommendations also called for increased variation in the methods used for assessing competences, as well as increased number of items, for improved construct validity.

The Expert Group's report was further contextualized by a concurrent curricular reform, enhancing focus on competences, critical reasoning skills, as well as explorative and in-depth learning. In addition, the new curriculum gave overarching educational aims like human dignity, identity and cultural awareness a place of privilege across all subjects, at the same time as it strengthened focus on interdisciplinary learning, through the prioritised themes democracy and citizenship, health and life skills and sustainable development.[7]

In addition to this important contextualizing factor, accessibility demands (WCAG) were also seen to be met most efficiently and fairly with the transition to the digital assessment platform. These three elements (Expert Group report, 2020 curricular reform, and WCAG) were therefore the central catalysts to the launch of the Norway's digitization project. The first two exams to be fully digitised in 2020 were English and maths. These two subjects were launched all-digitally first for private scheme candidates, owing to cancelled student exams because of the COVID-19 pandemic in 2021 and 2022. More subjects came online for the first full cohort examinations period of Spring, 2023.

[7] It may be worth noting here that the Directorate for Education was awarded a 2019 national prize for public transparency in official matters, for its management of the curricular renewal process. During this transition, thousands of subject experts, teachers and other education sector participants' comments were braided into the renewed curricular aims, resulting in a process that many felt satisfied with. On a personal note, it was this process of renewal that made me aware, as a lecturer in English didactics, of the strengths offered by a proactive and open Directorate. Shortly after participating in this process as an English subject expert from the outside, I applied for and was offered a position as English subject coordinator.

## New curricular coverage – increased validity, reliability

In secondary school English, the University of Bergen supported the development of the new assessment construct. Together with the Directorate's Examinations board, they looked carefully at the updated curriculum for 10th and 11th grades (the two subjects where examinations take place in English). They noted changes reflecting a marked shift from more of a social and cultural studies subject, to one that focused on aspects of using the English language, with some cultural aspects of English-speaking countries still included. Thus, the construct included a new "reception" section with listening and reading selected response items, in addition to the more traditional writing (production) sections.

These written production tasks, in turn, were updated as well. Two short, more "focused" writing tasks in mediation (target: 150 words) and interaction (target: 150-200 words) were included, as well as a more traditional longer form response (no target word count). In total, the five-hour written exam was given the following recommendation for time usage by pupils:

- Reception – 1 hour
- Mediation – 1 hour
- Interaction – 1 hour
- Longer written production – 2 hours

## New Methods

In addition to its new form, the exam in English also underwent a radical change in terms of methodology for its construction. Reception items were to be field trialed by pupils in the same year as those who would eventually take the exam. An analysis of the results would then be done using Item Response Theory (IRT), helping the Methods Team, Subject Lead and Examinations Board choose the items with the highest degree of construct validity. In this way, the exam's reliability was expected to improve, as the reception section (weighted as 1/3 of the total grade) would be marked automatically, relieving some of the pressure of written assessments' traditional challenges with inter-rater reliability.

The weighting of the exam was designed to balance the curricular aims and the time spent on the exam itself. With writing set at 2/3 of the total grade, productive skills are still very much emphasised. However, it is necessary to demonstrate both receptive and productive skills to pass the exam. This creates a situation where it is not possible to, say, skip the reception section, write for five hours, and still pass. This again comes back to construct validity, owing to the clear presence of listening and reading skills in the new curriculum.

Another new development in the assessment of the English exams was the establishment of a weighting system within the written task assessment criteria. It was determined, again based on the presence of so many language-learning competence aims, that language was to count for exactly half of the written task assessments. Ideas -- "content" -- would count for half as well. While separating written texts into such categories is a somewhat artificial process (occasionally written language is so full of non-standard usages that it becomes difficult to ascertain the intended meaning), for most pupils this represents an improvement over the traditional holistic approach that can make graders blind to good ideas because of “problematic language” usage.

This was in fact exactly what the Examinations Board, together with the team members from the University of Bergen found when the Directorate piloted a small-scale batch of written responses (N = 204), using our new, more analytical assessment criteria to assess them. We found that by separating out language from “content”, we were able to better discern the competences shown in responding to short and long response prompts, than if grading was done holistically. This small sample has been reconfirmed in several subsequent full-cohort examination settings, where co-graders report finding more agreement, more quickly during their common grading alignment sessions (“fellessensur”).

To support graders in their new, more analytical approach to grading, the Directorate developed a grading template, using spreadsheet development software. Here totals for Reception scores could be logged for hundreds of candidates along with candidate numbers, as well as sub-grades for relevance, independence and context. All these numbers have been properly weighted according to the prescribed weighting structure, without graders having to think about the maths involved. While spreadsheet software like Microsoft Excel is not a universally loved tool by English teachers, the strong majority of those who have encountered this new way of systematizing assessments has expressed gratitude for the guidance and overview it gives. Many have even adopted it for classroom use!

## New accommodations

One of the most challenging questions that came about as after the exam renewal was how to accommodate for pupils with hearing challenges. While the curriculum for pupils with sign language has its own parallel curriculum – with listening removed, obviously – listening skills in English were still a requirement (fairly or not) for pupils with challenges hearing.

After consultation with Statped, the national organ for accommodations in primary and secondary school learning and assessment, as well as families of pupils dependent on hearing accommodations for their learning, a field trial was devised in which videos with subtitles would be substituted for the listening files. These would be the new, accommodated “listening” tasks pupils with hearing challenges were tasked with answering. Field trials showed this solution was in fact a satisfactory accommodation that allowed pupils with hearing difficulties to demonstrate their receptive skills in English, using whatever sensory tools were at their disposal. Some pupils found listening possible, while others relied more on reading. Either way, the umbrella term of “reception” governed our conclusion that this was a fair and valid accommodation.

It should be noted that when the listening items were removed from the English exam for pupils with sign language, the Directorate together with the Examinations Board decided to add a reduced number of “perception” items. This term refers to the Norwegian verb “oppfatte”, used in the sign language curriculum instead of “listening”. For this competence aim, multimodal texts were designed which, because the reading load was going to be slightly heavier than for the common English exam, would be based on fewer texts, over fewer items compared to the listening tasks in the other English exam. This meant that for sign language pupils, the overall reading load would be minimized. At the same time, some important aspects of the test construction principles most suitable for IRT methods were maintained, such as using a large number of assessment items to establish test-takers’ receptive competence in English.

## Findings

While a larger study has yet to be published (expected in 2025-26), initial informal results show marked improvements in inter-rater reliability (IRR). This seems due to several factors including the exam's design, its weighting, and its use of more analytical assessment criteria.

In addition to some early quantitative calculations between co-graders' suggested grades that show the promise of significant IRR improvement, we have collected several semesters' worth of qualitative reports from grading training/calibration conferences and post-grading focus groups. These show nearly unanimously that graders experience improved assessment-interpretive communities, and more effective co-grading conferences, with higher rates of agreement, as well as more internal agreement on the sub-category grades in the wake of the digitization process.

We have also noted a largely unison student voice remarking positively on the changes the exam has undergone. Among the 300 responses to the example items we published in 2020, fully half came from pupils. It is probably not often that exams officials hear the word "cool" when referring to examination forms, but in this instance, we in fact received just that comment from a student who found it "cool" that listening was now part of the exam.

At a recent grading training conference, we found 90 percent agreement among the grading groups and the Examinations Board on six selected student exam responses. This eye-poppingly high rate of agreement is a figure more associated with mathematics assessment than language assessment. Perhaps unsurprisingly, graders left that conference with a high level of confidence that the assessment tools at their disposal were well suited to the challenges of assessing student writing.

It goes without saying that these improvements are not cost free. There remain challenges, including an increased workload, greater demands for quality assurance, foremost among them. Yet our shared experience, across the stakeholder horizon, has been that the strains are worth the improved process, and that for the time being, the results are satisfactory enough to tell us that the design changes are worth the trouble.

# Sources

Engeness, I. (2024). AI for Assessment for Learning. Østfold University College [Research project, ongoing]. https://www.hiof.no/lusp/pil/english/research/projects/ai4afl/index.html

Fullan, M., Azorín, C., Harris, A., & Jones, M. (2023). Artificial intelligence and school leadership: challenges, opportunities and implications. *School Leadership & Management*, DOI: 10.1080/13632434.2023.2246856

Garcia, B., Alario-Hoyos, C., Perez-Sanagustin, M., Morales, M., & Jerez, O. (2022). The Effects of the COVID-19 Pandemic on the Digital Competence of Educators. https://e-archivo.uc3m.es/rest/api/core/bitstreams/3fac6011-dffe-42a3-9dbd-ef799b86ed9e/content

Gerhardsen, M., Ed. (2024). Kunstig intelligens (KI) i Osloskolen. https://aktuelt.osloskolen.no/larerik-bruk-av-laringsteknologi/digital-skolehverdag/kunstig-intelligens-ki-i-osloskolen/

Gilje, (2024). Learning in the age of algorithms (LAA). [Research project, ongoing]. https://www.uv.uio.no/ils/english/research/projects/lat/index.html

Günther, J. H., Kløfte, R. M., & Hjorthen, I. R. (13 June, 2024). Skoler sliter med KI. Nå varsler regjeringen at de vil ta grep. (Schools are struggling with AI. Now the government is advising that it will take action. Author's translation). NRK.no. https://www.nrk.no/ostfold/skoler-sliter-med-ki.-na-varsler-regjeringen-at-de-vil-ta-grep-1.16883094

Miao, F., & Holmes, W. (2014). Guidance for generative AI in education and research. https://unesdoc.unesco.org/ark:/48223/pf0000386693.locale=zh

Molnes, G. (2024). Seks av ti lærere har tatt elever i KI-juks. (Six of ten teachers have caught pupils cheating). https://www.utdanningsnytt.no/chatgpt-juks-kunstig-intelligens/seks-av-ti-laerere-har-tatt-elever-i-ki-juks/386624

Nagel, I. M. L. (2024). Professional Digital Competence in Norwegian Teacher Education Policy and Practice: Teacher Educators' Professionalism in the Post-digital Age. https://www.duo.uio.no/handle/10852/107764

Norwegian Government. (1999). Lov om grunnskolen og den vidaregåande opplæringa (opplæringslova). https://lovdata.no/dokument/NL/lov/1998-07-17-61/KAPITTEL_1#KAPITTEL_1

Norwegian Research Council. (2024). Public Sector PhD Scheme. https://www.forskningsradet.no/en/apply-for-funding/funding-from-the-research-council/public-sector-phd-scheme/

Nelson, J., & Campbell, C. (2017). Evidence-informed practice in education: meanings and applications, *Educational Research*, 59:2, 127-135, DOI: 10.1080/00131881.2017.1314115

Osloskolen. (2024).

Rickinson, M., Cirkony, C., Walsh, L., Gleeson, L., Cutler, B., & Salisbury, M. (2022). A framework for understanding the quality of evidence use in education. *Educational Research*. 64:2, 133-158, DOI: 10.1080/00131881.2022.2054452

Røyne, H. (2024). Har innført «jukseverktøy» i undervisningen: – Stort potensial. (Have implemented "cheating tool" in instruction - great potential.) https://www.tv2.no/nyheter/innenriks/har-innfort-jukseverktoy-i-undervisningen-stort-potensial/16525563/

Russel Group. (4 July, 2023). New principles on use of AI in education. https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/

Utdanningsdirektoratet. (2023). Kompetansepakke om kunstig intelligens i skolen. (Competence package on AI in Schools). https://www.udir.no/kvalitet-og-kompetanse/digitalisering/kompetansepakke-om-kunstig-intelligens-i-skolen/#a194064

Utdanningsdirektoratet. (2024). Råd om kunstig intelligens i skolen. https://www.udir.no/kvalitet-og-kompetanse/digitalisering/kunstig-intelligens-ki-i-skolen/

Wasson, B. (2024). Artificial Intelligence in Education: Layers of Trust [Research project, ongoing]. https://slate.uib.no/projects/artificial-intelligence-in-education-layers-of-trust-edutrust-ai

# Serbia

# SERBIA: BIOGRAPHIES

## Katarina Aleksić

Katarina Aleksić is Head of Education Technology Center at the Institute for Education Quality and Evaluation. Her expertise centers on online and blended teaching and learning, with a particular emphasis on developing policies and strategies for quality digital education, as well as action planning for integrating digital education into both general and specific strategic documents. With a rich background as a computer science teacher, she has been honored with numerous international and national awards for her innovative teaching methods. Furthermore, Katarina has authored and co-authored several scientific papers and has contributed to eight textbooks in the areas of digital literacy and computer science.

## Katarina Glišić

Katarina Glišić is Advisor for Exam Programs Development and Test Instruments Preparation at the Institute for Education Quality and Evaluation in the Republic of Serbia. She has coordinated working groups in developing the Final Exam for compulsory education and has been actively involved in various training programs and educational projects. Additionally, Katarina is a licensed external school evaluator and has participated in numerous national and international educational conferences.

## Branislav Ranđelović

Branislav Randjelović, PhD is Director of the Institute for Education Quality and Evaluation of Republic of Serbia. He also works as Associate professor at Faculty of Electronic Engineering, University of Nis, and at Faculty of Teachers Education, University of K. Mitrovica. He published more than 100 scientific papers, research works, textbooks and publications.
He was researcher on various scientific and educational projects. Areas of research is applied mathematics, computers science, educational science and digitalization in education. He is official representative of Serbia in PISA governing board, TALIS governing board and IEA. He is engaged as NPM for PIRLS 2021, TALIS 2024 and PIRLS 2026.

# E-TESTING AND COMPUTER-BASED ASSESSMENT IN SERBIA[1]

## Abstract

This paper offers an overview of the achievements, current state, and trends in Serbian education concerning e-testing. It highlights examples from various educational segments and levels, including primary education (such as International Large Scale Assessments like TIMSS, PIRLS, and the Final Exam Field Trial), secondary education (including PISA), and e-testing in adult education. The Serbian educational system is highly focused on digitalization, aiming to ensure that all students develop digital competence. Current outcomes reflect progress toward this objective, and the entire educational system is eagerly awaiting the full implementation of e-testing procedures.

## INTRODUCTION

With advancements in digital technologies and their integration into educational systems, digital assessments are becoming increasingly important. Despite efforts by many education systems to regularly incorporate e-testing into the educational process, significant achievements remain limited. E-testing still appears to be in an experimental phase, with ongoing efforts accompanied by widespread skepticism.

Numerous researchers and scientists are contributing to this field. Notable studies include Al-Maawali et al. (2024), which explores teachers' perspectives on the affordances and challenges of technology for reliable and valid online testing, and Gunawan et al. (2024), which examines the design and validity of e-assessments. Haleem et al. (2022) focus on understanding the role of digital technologies in education, while Keskin et al. (2024) discuss the design of an assessment task analytics dashboard. Ortiz-Lopez et al. (2024) provide a mapping review of the role of e-assessment in the new digital context. Additional relevant literature includes Shute et al. (2017), which reviews computer-based assessment for learning in elementary and secondary education, and Weigand et al. (2024), which addresses mathematics teaching, learning, and assessment in the digital age. Vasu et al. (2024) explore assessment types and evaluation during the COVID-19 pandemic, and Kinnear et al. (2024) examine a collaboratively-derived research agenda for e-assessment in undergraduate mathematics.

In Serbia, research efforts include studies by Randjelovic et al. (2020) on online self-assessment as preparation for final exams in primary schools, and Randjelovic et al. (2022) on distance learning in Serbia and the experiences of primary education during the COVID-19 crisis.

[1] This work is partially funded This paper is supported by the Ministry of science and technological development through grants 451-03-65/2024-03/200102 and 451-03-65/2024-03/200251 and by the Ministry of Education through the internal project project 3.2. International assesments of Institute for Education Quality and Evaluation of Republic of Serbia.

The educational system of the Republic of Serbia is striving to align with contemporary issues and efforts, incorporating e-assessment and e-testing into various laws, regulations, and other relevant documents (Republic of Serbia, 2020; Republic of Serbia, 2023; Republic of Serbia, 2024).

However, e-testing and e-assessment have yet to become fully integrated into the Serbian educational system. Concerns persist regarding the availability of resources in schools, stable internet connections, and the potential lack of knowledge among teachers and students in using e-assessment tools. While there is a clear vision of what e-assessment should ideally look like, the current reality falls short of these expectations.

The structure of the paper consists of three chapters that cover various electronic assessments in Serbia over the past period. The next section explores how students in Serbia have used e-assessment tools during computer-based international large-scale assessments such as TIMSS, ICILS, and PISA (2022, 2023). The following section details the use of e-assessment tools in national assessments across various levels of primary education. Finally, the last section examines e-assessments in Serbia, particularly within adult education.

## COMPUTER-BASED TESTING IN INTERNATIONAL LARGE-SCALE ASSESSMENTS

Over the past two decades, the Serbian educational system has actively participated in numerous international large-scale assessments:

- Programme for International Student Assessment (PISA) since 2003,
- Trends in International Mathematics and Science Study (TIMSS) since 2003,
- Progress in International Reading Literacy Study (PIRLS) since 2021,
- International Civic and Citizenship Education Study (ICCS) since 2022,
- International Computer and Information Literacy Study (ICILS) since 2023,
- Teaching and Learning International Survey (TALIS) since 2024.

Serbia's engagement with computer-based assessments for 15-year-olds commenced with PISA 2018. During this assessment, test instruments were installed on computers, which were then transported to schools for student testing. The administration process was overseen by external researchers. Subsequently, in PISA 2022, instruments were distributed to schools via USB drives, and the administration process transitioned to school personnel. Looking ahead, PISA 2025 is anticipated to be conducted entirely online, with school staff continuing to manage the process.

For TIMSS 2023, the computer-based assessment for Grade 4 was conducted using USB instruments for the Field Trial in 2022 and fully online in the Main Survey in 2023. During the Field Trial, 16% of sampled schools reported a lack of appropriate digital devices for testing. In response, external laptops were provided by the IEQE for these schools. However, such intervention was unnecessary for the Main Survey, which was administered solely by school staff.

Regarding PIRLS 2026, the assessment is expected to transition to a fully online format, with school personnel handling the administration process.

During ICILS 2023, the computer-based assessment for Grade 8 students utilized USB instruments. After conducting compatibility checks, only 2.6% of sampled schools reported insufficient digital devices for testing. To address this issue, the IEQE provided external laptops to these schools as a solution. School staff managed the testing process.

The existing ICT infrastructure in Serbian primary schools proved sufficient for the successful implementation of computer-based assessments for the international large-scale assessments previously mentioned. This infrastructure, which includes desktop computers, laptops, internet access, and educational software, supported the administration of assessments such as PISA, TIMSS, and ICILS. Despite challenges, the existing infrastructure demonstrated its capability to support modern assessment methodologies. Teachers effectively integrated ICT into teaching, and professional development ensured their readiness for such assessments.

## COMPUTER-BASED TESTING IN NATIONAL ASSESSMENTS

The development of the Serbian eAssessment ecosystem has been gradual, with paper-based assessment still predominantly utilized in educational settings. Despite advancements in technology, there remains a prevailing reliance on traditional methods of testing. This slow progression towards eAssessment can be attributed to a lack of confidence in its fairness and credibility as a means of evaluating students' knowledge and skills. Stakeholders in the education system may harbor concerns about the reliability and validity of digital assessment methods compared to traditional paper-based exams. As a result, there is a hesitancy to fully embrace e-assessment as a primary means of evaluating student learning outcomes. This cautious approach underscores the need for ongoing efforts to address concerns and build trust in the effectiveness and integrity of digital assessment practices within the Serbian education system.

In addition to concerns regarding fairness and credibility, apprehension about potential malfunctions further contributes to the slow adoption of eAssessment in the Serbian education system. The fear of technical glitches or system failures during online assessments can undermine confidence in the reliability of digital testing platforms. Despite these reservations, Moodle has emerged as the preferred online testing platform for all large-scale eAssessments in Serbia. However, the persistence of these concerns highlights the need for robust infrastructure and comprehensive contingency plans to address technical challenges and ensure the smooth implementation of eAssessment initiatives. Efforts to mitigate the risk of malfunctioning and enhance the resilience of digital assessment systems are essential for fostering greater confidence in the transition towards eAssessment within the Serbian educational landscape.

## *Final Exam and Adapting Primary Education to Changing Circumstances*

Primary education in Serbia typically spans eight years, starting at the age of six or seven. It is divided into two stages: the First cycle (grades 1-4) and the Second cycle (grades 5-8). The curriculum covers a range of subjects including mathematics, language arts, science, social studies, physical education, and arts. Primary education aims to provide students with a solid foundation of knowledge and skills while fostering critical thinking and social development.

To obtain a primary school diploma in Serbia, students typically need to fulfill these requirements:

- Completion of eight years of primary education, covering grades 1-8 with satisfactory attendance and performance and
- Successful completion of the Final Exam administered by educational authorities.

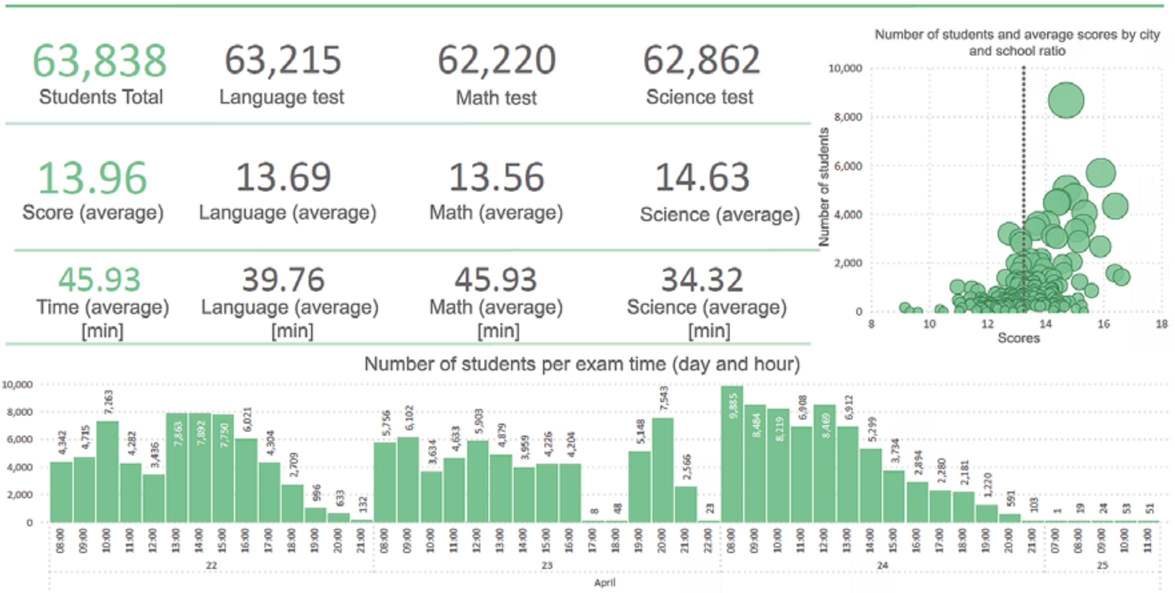
The concept of the Final Exam[2] was established in 2010, aiming to provide data on the level of achievement of general and specific standards, namely educational standards for the end of compulsory education. This serves as the basis for evaluating the quality of compulsory education. In developing the concept of the exam, three main functions are considered: certification, selection, and evaluation. After completing the Final Exam, the student is considered to have finished Primary Education and gained the right to enroll in secondary school, which means that there is no minimum knowledge required for this exam.

The tests in the Final Exam contain tasks that assess the achievement of educational standards for the end of compulsory education for students who have completed primary school. Students take the final exam by solving tasks within three tests: in the subject of Serbian language and literature, or mother tongue and literature; in the subject of Mathematics; and one of the five subjects chosen by the student from the list of subjects from natural and social sciences: Biology, Geography, History, Physics, and Chemistry. The tests in the final exam contain tasks that assess the achievement of educational standards defined at three levels of attainment – basic, intermediate, and advanced. These levels describe requirements of varying difficulty, cognitive complexity, and extent of knowledge, ranging from simple to complex.

The Final Exam is a comprehensive assessment taken by all eighth-grade students completing primary education. It takes place in June. Before the final exam, the Ministry of Education conducts a Final Exam Field Trial (FE Field Trial) to prepare students for all aspects they will encounter during the exam. This trial helps students understand the time frame needed for their work, the procedures involved, and how their achievements will be assessed. Furthermore, the FE Field Trial offers students a chance to identify curriculum areas they need to concentrate on. If they find uncertainty in specific areas, they can plan their further study more effectively.

[2] All provisions related to the final exam are established by the Law on the Foundations of the Education System („Official Gazette of RS“, no. 88/17, 27/18 - other law, 10/19, 6/20, and 129/21), the Law on Primary Education and Upbringing („Official Gazette of RS“, no. 55/13, 101/17, 27/18 - other law, 10/19, and 129/21), and the Regulation on the Program of the Final Exam in Primary Education and Upbringing („Official Gazette of RS – Educational Gazette“ no. 1/11, 1/12, 1/14, 12/14, 2/18, 3/21, and 14/22).

In 2020, the Serbian education system faced significant challenges during the COVID-19 pandemic. Initially, schools closed in March 2020, transitioning to online learning. However, disparities in access to technology and internet connectivity among students posed obstacles. The government attempted to address these issues by providing online learning platforms and distributing educational materials. As the situation evolved, there were intermittent periods of in-person classes with safety measures in place. Ultimately, the response varied across regions and educational levels, with efforts to balance educational continuity and health concerns. Since Final Exam is annual activity and FE Field Trial is an important part of it, the Ministry of Education, Science and Technpological Development organized a new form of this activity – *FE Field Trial Online Self-evaluation* (Randjelovic et al, 2020). This was the largest (Fig. 1) e-assessment endeavor in the Serbian Education System.



63,838 Students Total

63,215 Language test

62,220 Math test

62,862 Science test

13.96 Score (average)

13.69 Language (average)

13.56 Math (average)

14.63 Science (average)

45.93 Time (average) [min]

39.76 Language (average) [min]

45.93 Math (average) [min]

34.32 Science (average) [min]

*Fig. 1 Statistic overview on FE Field Trial Online Self-evaluation 2020*

The *FE Field Trial Online Self-evaluation* was designed to allow students to evaluate their knowledge by solving three tests covering a total of seven subjects (Serbian language, Mathematics, Physics, Geography, History, Chemistry, and Biology) two months before taking the Final Exam at the end of elementary education. This approach facilitated the development of self-regulated learning, primarily through self-assessment of current knowledge levels and planning subsequent learning steps.

In situations where schooling is conducted remotely, the *FE Field Trial Online Self-evaluation* tests, provided through the joint efforts of the Ministry of Education, Science, and Technological Development, the Institute for Education Quality and Evaluation, the Office for IT and Electronic Administration, and Comtrade Company, offered additional support to students, teachers, and schools via the online Moodle platform. From April 22 to April 23, out of a total of 68,504 eighth-grade students, 63,215 solved the Serbian language test, 62,200 Mathematics, and 62,825 the Combined test, indicating that over 91% of students took each test. It was evident that many students independently solved the tests and used the FE Field Trial 2020 in the best possible way for future academic growth.

The examinations took place on Comtrade Company's Moodle platform, supervised by a ten-member IT team. Technical challenges emerged during the testing period, especially when a substantial number of students accessed the platform concurrently. These issues were promptly managed, primarily through enhancing hardware capabilities and subsequently implementing an hourly access rate for tests in different regions. Notably, there were high public expectations for flawless execution, despite Serbia having no prior experience with such a large-scale online knowledge assessment.

The "Final Exam - Field Trial 2020" marked a historic milestone in Serbian education as the first online assessment for an entire generation. Although optimism surrounded the potential for this practice to become standard, its repetition has yet to occur, underscoring the ongoing evolution and adaptation of educational practices in response to changing circumstances.

## *Digital Competence Assessment at the end of Primary Education in 2016*

In Serbia, Computer Science was integrated into the mandatory curriculum for students aged 11-14 as part of the second cycle of primary education. This initiative was launched in the academic year 2017/18, involving nearly 250,000 students in exploring various aspects of Computer Science.

Before Computer Science became a mandatory subject, students acquired digital competencies through two school subjects: Technical and Informatics Education (compulsory for all students, with limited digital content) and Computer Science (an elective subject focusing solely on digital competencies, attended by a subset of students).

The introduction of Computer Science as a mandatory subject stemmed from a national computer-based assessment conducted in 2016 to gauge students’ digital competence after primary education. Led by the Ministry of Education, Science, and Technological Development of the Republic of Serbia, with support from the Institute of Psychology at the Faculty of Philosophy in Belgrade, the assessment collected data from a sample of 56 primary schools across Serbia. This sample, stratified by region and locality size, comprised 1014 8th-grade students, of which 949 provided complete responses. The research was conducted online on May 5, 2016, via the Moodle platform, lasting approximately 45 minutes.

The research examined students’ use of ICT (computers, mobile phones, internet) both within and outside the school environment, their understanding of basic computer concepts, and the application of this knowledge in real problem-solving situations. It covered various areas such as hardware, operating systems, office programs, graphics, word processors, presentation software, internet, security, ethics, online violence, programming, mobile phones, and spreadsheet software. The knowledge test consisted of 30 problem-solving tasks set in real-life contexts, presented as multiple-choice questions. Additionally, a questionnaire gathered information on students’ socio-demographic characteristics, ICT usage, activities during ICT classes, and attitudes towards computers and internet usage. Testing was conducted on the Moodle platform provided by the Ministry of Education, Science, and Technological Development, with no technical issues reported.

Student performance on the knowledge test ranged from 2 to 27 points out of a possible 30, with a mean score of 12.80 points. The test results indicated varying levels of digital competence among students, with correlations observed between test performance and factors such as overall academic performance, self-assessment of computer skills, time spent online, enrollment in elective courses in Computer Science, and parental education levels. However, certain factors, such as gender differences, frequency of ICT use in school, and students' perception of task difficulty, showed negative correlations with test performance.

The findings from the 2016 questionnaire revealed that students primarily used digital devices and the Internet for entertainment purposes outside of school. Additionally, a significant proportion of students did not use learning programs at home or by email. Despite high online activity, only a small percentage of students attended elective courses on Computer Science or utilized learning platforms at school.

In 2016, the Ministry of Education, Science, and Technological Development conducted an e-assessment of students' digital competencies at the end of primary education. The research data collected served as a foundational dataset for introducing Computer Science as a mandatory subject in Serbian primary schools. By systematically examining students' proficiency levels, ICT usage patterns, and attitudes toward technology, this study provided crucial insights into the existing landscape of digital education. Decision-makers recognized the imperative for evidence-based policies, understanding that educational reforms must be grounded in comprehensive data analysis to effectively address the evolving needs of students in the digital age. Consequently, the integration of Computer Science into the curriculum was a strategic response to the findings of this research, aiming to equip students with essential skills for navigating an increasingly technology-driven society. This exemplifies the importance of utilizing empirical research as a cornerstone for shaping educational policies and initiatives.

## *Digital competence of 4th-grade students in primary schools*

The examination of the level of digital competence among fourth-grade primary school students is part of the Quality Assurance of Digital Technology Integration in the Education System of the Republic of Serbia project, implemented by the Institute for the Evaluation Quality and Evaluation (hereafter referred to as IEQE) – Education Technology Center. One of the main objectives of this project, as well as one of the strategic goals of educational policymakers in recent years, is to improve the Quality Assurance System of Digital Education in the Republic of Serbia.

To develop and enhance students' digital competencies, over the past four school years (from 2020/21 to 2023/24), the subject Digital World has gradually been introduced as a mandatory subject for all younger primary school students in the Republic of Serbia. The curriculum of the Digital World subject is designed to allow students to acquire appropriate knowledge, skills, and attitudes during 36 hours per year, enabling them to use digital devices safely and effectively for learning, communication, and collaboration, both in the school and out-of-school contexts. Within this subject, students should also acquire the fundamentals of algorithmic thinking as a prerequisite for acquiring knowledge and skills in programming (during the later grades of primary and secondary school), as well as for successful navigation in everyday activities in today's deeply digitized society.

Using the same online tool, under identical technical and organizational conditions, two studies were conducted:

- Baseline research (autumn 2022/23) in which the level of digital competence of fourth-grade students who had never attended the Digital World subject was directly examined, and
- Main research (autumn 2023/24) in which the level of digital competence of fourth-grade students who had studied the compulsory subject Digital World since the 2020/21 school year was directly examined.

The purpose of these studies was to determine and compare the current level of development of digital competencies among students of this age group, primarily considering assessing the contribution of studying the Digital World subject to the students' actual digital competence level. Given that the same online assessment tool was used in the studies, containing tasks thematically related to the teaching content and prescribed learning outcomes of the Digital World subject, it was realistic to expect that the test achievements of students who had attended this compulsory subject would be higher compared to those who had not.

This report represents a comparative analysis of empirical data and an integration of findings obtained from two online tests of digital competencies of fourth-grade students completing the first cycle of primary education in the Republic of Serbia.

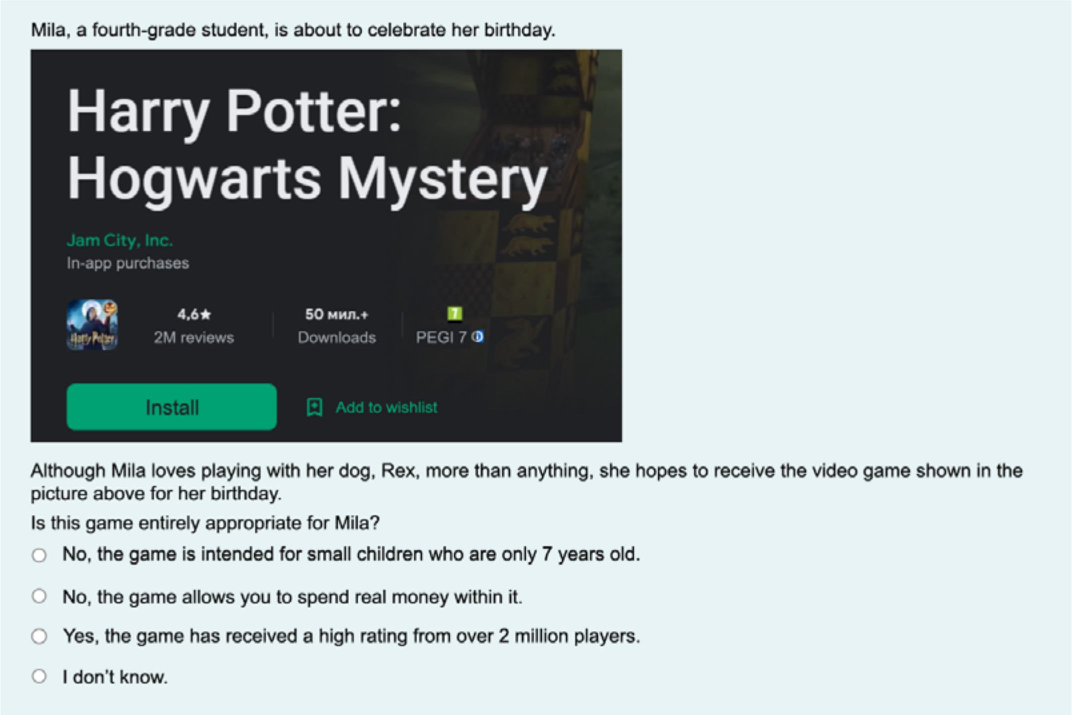
The online tests were conducted on the IEQE platform, within the Moodle learning management system. The research was carried out by trained school coordinators based on detailed instructions received in advance (it was recommended to consider the level of their digital competencies when selecting school coordinators). The maximum time for working on the test was 60 minutes. After the time expired, the test would automatically close (students could finish the test earlier, but after submitting the test, they could not change their answers). School coordinators did not report any technical problems during the testing.

The sample of students (in both studies) was random and stratified by region and locality size. Various schools were included in the baseline and main research, but attention was paid to geographical location and locality size (in this regard, the samples did not differ). Schools participating in the main research were selected to be from the same municipalities and cities as the schools participating in the baseline research. In the case of rural schools, if there was no school from the same village as in the baseline research, a school from a similar village in the same municipality was chosen. This selection method provided a good basis for further comparative analysis of the obtained data. IEQE did not collect any personal data of students. Schools only reported the number of students in the first class of fourth grade, after which they received usernames and passwords from IEQE for each student to access the online test. For two schools that were not adequately equipped (they were in the process of renovating computer labs), computers were provided for testing, which is why the research was not conducted at the same time in all schools, but it was in most schools included in the sample.

A total of 1004 fourth-grade students from 55 primary schools in Serbia participated in the online test of the baseline research, but 986 students were included in the statistical analysis[3]. A total of 979 fourth-grade students from 56 primary schools in Serbia participated in the online test of the main research, but 947 students were included in the statistical analysis[4].

For the research, an online tool was created – a test for direct assessment of the digital competencies of fourth-grade primary school students. Closed-type tasks created for this research were carefully designed to examine the level of achievement of selected learning outcomes of the Digital World subject. This approach enabled the identification of areas where students show the most progress, as well as areas where additional efforts are needed to achieve defined learning outcomes.

The test questions were designed so that the prescribed outcomes of the Digital World subject were "embedded" in real-life contexts close to students' experiences (see Fig. 2). This allowed insight into how students who had never attended the Digital World subject, as well as those who had, cope with situations reflecting the real world they live in. The created test questions were grounded in 24 outcomes measurable in an automated knowledge assessment situation in an online environment. It is important to note that the test tasks from the Computational Thinking area were of moderate complexity and checked cognitively simpler outcomes precisely because of the participants of the baseline research – students who had no formal education in the field of informatics and computer science.


Mila, a fourth-grade student, is about to celebrate her birthday.

Although Mila loves playing with her dog, Rex, more than anything, she hopes to receive the video game shown in the picture above for her birthday.

Is this game entirely appropriate for Mila?

○ No, the game is intended for small children who are only 7 years old.

○ No, the game allows you to spend real money within it.

○ Yes, the game has received a high rating from over 2 million players.

○ I don't know.

Fig. 2 Test question designed to "embedd" Digital World in real-life contexts

[3] The following 18 students were excluded from the analysis: one student who did not do any of the tasks in the test and students who (judging by the time spent) answered mechanically (five students completed the work on the test in less than 2 minutes, the remaining 12 students in less than 10 minutes).

[4] A total of 32 students were excluded from the analysis: students who did not do any of the tasks (28 of them), more precisely, students who either did not submit the test or did not click on the End attempt button (the reason is unknown - maybe they were working on tasks, maybe they had problems of a technical or other nature) and 4 students who completed the test in about 6-7 minutes, and they scored between 9 and 18 points (during this time the students could hardly read all the tasks).

Specifically, the test consisted of 21 closed-type questions, with four options provided, one of which was correct. The questions were presented to all students in the same order, but the arrangement of answer options (two distractors, the correct answer, and the answer I don't know) was different (random). Seven tasks/questions from the test related to the teaching area Digital Society, eight to the teaching area Safe Use of Digital Devices, and six to the teaching area Computational Thinking.

As mentioned above, the research was conducted to determine the pedagogical benefits of attending the Digital World school subject. Specifically, it was expected that students attending this subject in the fourth grade would achieve higher scores on the digital competence test.

Regarding the overall achievement of students on the digital competence test, a normal distribution was obtained, which was slightly "shifted" to the left in the baseline research, indicating that the test was slightly more challenging for students who did not attend the Digital World subject, as well as for students who did attend this subject. Although the average student achievement on the test changed very slightly, averaging 9.3 in the baseline research and 10.3 in the main research, a statistically significant difference in favor of students who attended the Digital World subject was found at the whole sample level. It should be noted that statistical significance does not necessarily imply the practical significance of the observed difference.

Based on the findings, several key recommendations have been formulated to address the observed challenges and ensure the attainment of the prescribed learning outcomes for the Digital World subject:

1. Research the attitudes and specific practices of primary school teachers related to the implementation of the mandatory subject Digital World.
2. Conduct research on the level of satisfaction of students and their parents with learning and teaching in the Digital World subject.
3. Develop new highly effective professional development programs and other forms of professional support for primary school teachers, particularly in the areas of algorithmic thinking and programming.
4. Strengthen professional pedagogical supervision of the implementation of the Digital World subject within school administrations.
5. During the external evaluation process of primary schools, it is essential to visit Digital World school hours.
6. Determine the level of infrastructure readiness of schools and the availability of appropriate digital resources for teachers and students in the younger grades of primary school.
7. Determine the level of participation of primary school teachers in existing training sessions for the implementation of the Digital World subject I-IV conducted by the Institute for the Improvement of Education and the level of use of digital teaching materials made available to teachers within these trainings.
8. Further enhance the online tool for the direct assessment of the digital competence level of fourth-grade primary school students (e.g., increase the number of items for each of the teaching areas, include data on student grades, gender, etc.).

Given that the research results were published in 2024, it is still premature to discuss the implications arising from them.

## COMPUTER-BASED TESTING IN ADULT EDUCATION

From 2020, tree large-scale assessments of digital competence of adults in Serbia were conducted by the IEQE on an online Moodle platform:

- *Upskilling Pathways: New Opportunities for Adult Work Skills Development 2020.*
- *Serbia at your fingertips – Digital Transformation for development 2020.*
- *New Skills for Emerging Industries - National IT Retraining Programme 2022.*

### *Upskilling Pathways: New Opportunities for Adult Work Skills Development 2020*

The *Upskilling Pathways: New Opportunities for Adult Work Skills Development* project is being executed in the Republic of Serbia under the EU Program for Employment and Social Innovation. As per the project contract, the IEQE has delineated minimum requisite skills in linguistic, numerical, and digital literacy essential for successful employment. These standards were sourced from relevant domestic and international legal documents[5], as well as from findings of pertinent international research listed in the Literature section at the end of this paper.

According to the 2011 census data, nearly 35% of individuals aged 15 and above in Serbia have completed primary education or lower. Among them, 2.68% have received no formal education, and 11% have incomplete primary education. These statistics underscore the importance of establishing proficiency standards for individuals entering the workforce. IEQE identified a total of 58 descriptors outlining minimum language, numerical, and digital literacy requirements for work and daily life in Serbia. Among these, 27 descriptors were selected for validation, while others, though significant, were not tested due to constraints such as test duration or the necessity of specific devices, primarily computers.

The validation testing included 1,829 individuals classified as low-skilled unemployed persons registered with the National Employment Service. The results show that the average score across the entire assessment was 14.7 out of a possible 30 points, indicating that, on average, the tested population only reaches half of the recommended minimum proficiency levels in language, numerical, and digital literacy. Additionally, none of the selected skills or abilities were fully mastered by all respondents. Following these findings, the IEQE has developed

[5] In 2006, the European Parliament and the Council of the European Union adopted the Recommendation on Key Competences for Lifelong Learning. In May 2018, the Council adopted a recommendation for a new framework of key competences for lifelong learning. The "European Reference Framework of Key Competences" defines the competencies needed by every individual for personal fulfillment and development, employment, social and civic inclusion.

The Republic of Serbia, in accordance with its efforts to join the European Union as a full member, has defined, on the basis of the aforementioned documents, "key competencies as a set of integrated knowledge, skills and attitudes that each individual needs for personal fulfillment and development, inclusion in social life and employment" ( Law on the Basics of Education and Training, "Official Gazette of RS", No. 88/2017 of 29.9.2017).

proposals for the content and format of the training programs. The overall conclusion is that without a defined, adopted, and widely promoted minimum set of linguistic, numerical, and digital competencies essential for work and daily life, the creation of training programs for unemployed individuals with low or no qualifications is at risk of being based on the subjective judgments of the organizers.

Ensuring a high quality of life and professional competency in modern society requires meeting specific knowledge standards and developing diverse skills. Individuals must navigate work environments adeptly, address challenges effectively, and utilize resources and technology responsibly. Given that the newly defined standards reflect the outcomes of both formal education and other forms of learning (informal learning outside educational institutions and extra-institutional learning in everyday life), they serve as a necessary starting point for defining relevant retraining programs for hard-to-employ citizens.

### *Serbia at your fingertips – Digital Transformation for development 2020*

As a candidate for EU membership, one of Serbia's key objectives is to join the EU and its single (digital) market. Achieving this goal requires strengthening the capacities of the Serbian economy and administration. Digital transformation is a top government priority, and Serbia aims to swiftly implement efficient, secure, and citizen-oriented e-services, as well as to coordinate ICT policy implementation. Serbia at your fingertips – digital transformation for development project aimed to prepare and support the Serbian public administration and economy for digital transformation and to enable the Government of Serbia to provide more transparent and accountable digital services that meet the expectations of citizens and the needs of the economy. One of the key activities was strengthening the IT sector in the Republic of Serbia by addressing the workforce shortage through informal education in the skills most demanded by the industry.

Institute for Education Quality and Evaluation developed and implemented online testing for candidates who wished to participate in IT retraining programs. The goal was a high-quality selection of candidates. The areas of work for IEQE in this project included:

- Developing a testing methodology that can effectively assess logical and algorithmic thinking, English language proficiency, and other relevant aspects for participation in the IT retraining program.
- Conducting online testing for a wide range of candidates.
- Creating an objective ranking of all candidates based on their performance and demonstrated knowledge on the test.

For this project task, the project team created a question bank on the Moodle platform. The bank contained:

- 102 questions for the personality test,
- 30 questions-tasks for the intelligence test,
- A test with 60 questions on algorithmic thinking, logical reasoning, and problem-solving,
- A test with 20 questions on English language proficiency.

It was crucial to properly set the duration of the course/testing. It was also essential to align the duration of the course/testing with the difficulty of the prepared questions. All these adjustments were necessary to ensure the relevance of the results and to carry out the candidate selection in the best possible manner.

The online testing was conducted in November 2020 when 2,629 candidates participated. Minor problems related to the Cron job occurred and were addressed immediately.

### *New Skills for Emerging Industries - National IT Retraining Programme 2022*

In cooperation with UNDP Serbia, the Institute for Education Quality and Evaluation developed and implemented online testing for candidates interested in IT retraining programs. The objective of this project activity was to facilitate a meticulous selection process for participants of the IT retraining program. This program aimed to bolster the IT sector in the Republic of Serbia by addressing the shortage of skilled professionals through informal education in high-demand industry skills.

The platform received registrations from 1,491 candidates interested in participating in the IT sector retraining program. It was designed to offer flexibility, allowing candidates to participate in activities at their convenience during the 9-day testing period. Each test was accessible only once. Upon completion, candidates could submit the test independently by clicking a designated button. If a candidate failed to submit the test manually, the platform automatically did so. Additionally, if a candidate exceeded the allotted time for the test, it was automatically submitted.

Although the number of registered candidates did not pose a threat to the functionality of the IEQE Moodle platform, it was decided to suspend all other Institute activities during the testing period. Additionally, the Cronjob activity was set to activate every minute to ensure smooth operation. The online testing was conducted in June 2022.

## CONCLUSION

In conclusion, this paper offers a comprehensive overview of the achievements, current status, and emerging trends in e-testing within Serbian education. We have illustrated the progress made through various examples from different educational levels, including primary (International Large Scale Assessments such as TIMSS, PIRLS, and final exams), secondary (PISA), and adult education. The Serbian educational system is firmly committed to digitalization, aiming to equip all students with the necessary digital competencies. The results to date reflect this commitment, with the system eagerly anticipating the full implementation of digital testing procedures.

International Large Scale Assessments (TIMSS, PIRLS, PISA, ICCS, ICILS, TALIS) are increasingly adopting digital formats, a transition that has been notably accelerated by the pandemic. To stay aligned with global standards and ensure comparability of student performance, our educational system must continue to adapt to these changes. The successes and results of these assessments confirm that Serbian students are keeping pace with technological advancements and are well-prepared for the global labor market.

National assessments, particularly the final exams at the end of primary school, are essential for advancing digital literacy among students, which is critical for the future of our society.

wAssessments of digital competencies for students (ICILS and national assessments), teachers (TALIS and national questionnaires), and parents (integrated with national and international assessments) underscore the significance of digital skills across all life domains, not just within education.

This paper marks the beginning of a broader discussion. The integration of e-assessment and e-testing into the Serbian educational system is expected to grow more robust, especially with the advent of AI tools, which will introduce new challenges for both educators and learners. Insights from research by Farazouli et al. (2024) and Siddiq et al. (2024) provide an early glimpse into these future developments.

Overall, while we are at the start of this journey, it is evident that the digital transformation in education is both inevitable and necessary.

## REFERENCES

Al-Maawali, W., & Al Rushaidi, I. (2024). Teachers' Perspectives on Affordances and Challenges of Technology for Reliable and Valid Online Testing: Learning the Lessons from the COVID-19 Pandemic. Ubiquitous Learning: An International Journal, 17(1).

Farazouli, A., Cerratto-Pargman, T., Bolander-Laksov, K., & McGrath, C. (2024). Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practices. Assessment & Evaluation in Higher Education, 49(3), 363-375.

Gunawan, M. S., & Mufit, F. (2024). DESIGN AND VALIDITY OF AN E-ASSESSMENT FOUR TIER MULTIPLE CHOICE TO ASSESS SCIENCE LITERACY SKILLS OF SENIOR HIGH SCHOOL STUDENTS. Physics Learning and Education, 2(1), 28-38.

Haleem et al., (2022) Understanding the role of digital technologies in education: A review, Sustainable Operations and Computers,Volume 3, Pages 275-285, ISSN 2666-4127, https://doi.org/10.1016/j.susoc.2022.05.004

Keskin, S., Aydın, F., & Yurdugül, H. (2024). Design of Assessment Task Analytics Dashboard Based on Elo Rating in E-Assessment. In Assessment Analytics in Education: Designs, Methods and Solutions (pp. 173-188). Cham: Springer International Publishing.

Kinnear,G.,Jones,I.,Sangwin,C.,Alarfaj,M.,Davies,B.,Fearn,S.,...&Wong,T.(2024).Acollaboratively-derived research agenda for e-assessment in undergraduate mathematics. International Journal of Research in Undergraduate Mathematics Education, 10(1), 201-231.

Ortiz-Lopez, A., Olmos-Miguelanez, S., & Sanchez-Prieto, J. C. (2024). Toward a new educational reality: A mapping review of the role of e-assessment in the new digital context. Education and Information Technologies, 29(6), 7053-7080.

Randjelovic, B., Aleksic, K., Stanojevic, D. (2020). Online self-assesment as preparation for final exam in primary schools – experience from COVID19 crisis, Proceedings of **International conference on Applied Internet and Information Technologies** AIIT2020, Technical faculty Zrenjanin, p.297-300.

Randjelovic, B., Karalic, E., Djukic, D. (2020). The Digitalization of the Learning Process in Serbia During COVID-19 Crisis, Proceedings of Int.Sci.Conf. "Nauka i nastava u vaspitno-obrazovnom kontektstu", Pedagoski fakultet Uzice pp.203-216. http://doi.org/10.46793/STEC20.203R

Randjelovic, B., Karalic E., Djukic. D., Aleksic, K. (2022). Distance Learning in Serbia - Experience in Primary Education during COVID-19 Crisis, TEME, Vol. XLVI, 2, 377-397. https://doi.org/10.22190/TEME210609024R

Republic of Serbia (2023), Zakon o osnovnom obrazovanju i vaspitanju, "Official Gazette Republic of Serbia", 92/2023.

Republic of Serbia (2020), Strategija razvoja digitalnih veština u RS za period od 2020. do 2024. godine, "Official Gazette Republic of Serbia", 12/2020, 8/2023.

Republic of Serbia (2023), Akcioni plan za sprovođenje Strategije razvoja digitalnih veština u Republici Srbiji za period od 2020. do 2024. godine, u periodu od 2023. do 2024. godine, , "Official Gazette Republic of Serbia" 8/2023.

Republic of Serbia (2024), Pravilnik o bližim uslovima za sprovođenje državnih i međunarodnih testiranja, "Official Gazette Republic of Serbia", 34/2024.

Shute, V. J. & Rahimi. S. (2017). Review of computer-based assessment for learning in elementary and secondary education. Journal of computer assisted learning. Vol 33(1), p. 1–19. https://doi.org/10.1111/jcal.12172

Siddiq, F., Olofsson, A. D., Lindberg, J. O., & Tomczyk, L. (2024). What will be the new normal? Digital competence and 21st-century skills: critical and emergent issues in education. Education and Information Technologies, 29(6), 7697-7705.

Vasu, K., Supian, N., Nimehchisalem, V., Kirshner, J. D., & Ling, S. C. Y. (2024). A Review of Types of Assessment and Evaluation during the COVID-19 Pandemic across the Continents (General Assessment). Creative Education, 15(6), 1125-1139.

Weigand, H. G., Trgalova, J., & Tabach, M. (2024). Mathematics teaching, learning, and assessment in the digital age. ZDM–Mathematics Education, 1-17.

# Slovenia

# SLOVENIA: BIOGRAPHIES

## Ana Radović

Ana Radović holds degrees in English and French language and literature. She worked as a teacher in language and primary schools before joining the National Examinations Centre in 2010, where she has since worked as an examinations manager. Her responsibilities include overseeing language examinations at the National Assessment in primary education, coordinating item-writing groups, and providing administrative and expert support. She participates in test development and e-marking procedures. Ana has been involved in various projects of the National Examinations Centre, such as the implementation of e-marking, a pilot e-testing project, linking foreign language examinations to the CEFR, etc. She is especially interested in questions concerning the future of assessment.

## Andrejka Slavec Gornik

Andrejka Slavec Gornik holds degrees in Geography and Sociology of Culture, and a PhD in Geography. She worked as a young researcher and assistant at the Department of Geography of the Faculty of Arts in Ljubljana for seven years before joining the National Examinations Centre in 1996. Initially, she worked as an examinations manager and later became the Head of the Examinations Unit. In her role, she oversees and coordinates the work of examinations managers and testing committees for the General and Vocational Matura examinations, as well as for the National Assessment in primary education. She prepares expert documents and has an advisory role in various national committees. She has conducted numerous seminars on test construction, examination syllabus preparation, quality assurance of external marking, etc. Andrejka has been involved in various international projects, including the introduction of the Matura examination in Bosnia and Herzegovina and in Serbia.

# SLOVENIA'S TRANSITION TO E-MARKING AT NATIONAL EXAMINATIONS

## Abstract

Slovenia's National Examinations Centre administers external assessments, including the National Assessment at the end of grades 6 and 9 in primary education and Matura examinations at the end of secondary education. In recent years, we have completed the transition from paper-based marking to electronic marking of examinations at both levels. The article describes Slovenia's experience with the implementation of e-marking, the activities that preceded the introduction of e-marking, the challenges that we faced during the preparation and implementation phases, and the benefits that resulted from the implementation of e-marking. A pilot e-testing project that we administered in 2021 will also be briefly addressed, as well as the challenges that lie ahead.

## 1. Introduction

The National Examinations Centre (NEC) is the central examination board in Slovenia, responsible for administering different types of national examinations, marking them, distributing results, and awarding qualifications.

In primary education, each year approximately 20,000 sixth-graders (11- to 12-year-olds) and around 20,000 ninth-graders (14- to 15-year-olds) take the National Assessment. The National Assessment at the end of grade 6 consists of examinations in L1 (Slovenian, Italian and Hungarian), mathematics, and L2 (English and German). At the end of grade 9, students are tested in L1 and mathematics again, while the third subject is determined each year by the Minister of Education.

At the end of secondary education, students (18- to 19-year-olds) take either General Matura (about 6,000 each year) or Vocational Matura (around 10,000 each year). At General Matura, each candidate sits exams in five subjects, while at Vocational Matura, candidates take exams in four subjects.

As a whole, this amounts to about 180,000 scripts[1] that need to be marked in a two-month spring session (only a small percentage of scripts fall into the autumn session).

Before the year 2013, all scripts were marked manually. By then, we had identified numerous problems connected with paper-based marking. Besides organizational, security and quality issues, we also noticed reluctance on the part of examiners to do "double work" (marking students' scripts and then transferring their marks on scanning sheets) and technical errors

[1] Scripts are individual exam papers written by students in an exam.

such as miscalculations. Our need to tackle these problems, combined with a wish to modernize the assessment process, coincided with the political and financial support that we received from decision-makers at the Ministry of Education in 2011. We therefore decided that in the school year 2012/2013, the scripts of ninth graders at the National Assessment would be electronically marked for the first time, and a year later, the same form of marking would be used at the National Assessment in grade 6. In the years that followed, we gradually introduced e-marking at General Matura examinations as well.

Today, all scripts that we get at the NEC are marked online. The exception is the Vocational Matura, where the scripts stay at schools and are still marked manually. For the National Assessment and General Matura examinations, we have successfully completed the transition from paper-based to onscreen marking.

In this article, we will explain how we prepared for the introduction of e-marking in a short period of time, how we organized the training network, what problems we encountered during the preparation and implementation stages, and what responses we got from the parties involved. We will also try to present the overall effects that the implementation of e-marking had on the organization, security and quality of marking procedures. The Slovenian case study will be of interest primarily to those who are planning a similar project in the future and will find our experience with the implementation of e-marking useful.

## 2. Prior experience with e-marking (Pilot project in 2009)

Before introducing e-marking at the national level for the first time in 2013, the NEC had been striving to make necessary changes in the marking system for several years, especially for the General Matura exams. The General Matura in Slovenia is a high-stakes examination and is required for the completion of secondary education. The Matura results are also used for selection purposes for university admissions. Therefore, the quality of external marking is extremely important.

With an awareness that with e-marking we could enhance the quality of marking, the NEC decided to carry out a pilot e-marking project together with the British company RM Education in 2009. The main aims and objectives of the project were as follows:

- To evaluate the quality and suitability of the chosen e-marking software.
- To get familiar with the technical prerequisites for the introduction of e-marking (preparation and digitalization of material, database preparation, script allocation, control over the marking process, etc.).
- To analyze the impact of new technology on the quality of marking.
- To analyze the acceptance of new technology by external examiners and gain their opinion on it.
- To analyze the quality of e-marking in comparison with the existing quality of paper-based marking.

- To carry out a study to determine whether the software is suitable for the introduction of e-marking in all General Matura subjects.
- To estimate the costs and savings that e-marking would bring.

(Slavec Gornik & Urank, 2009)

During the pilot project, one paper of each exam in compulsory General Matura subjects (Slovenian, mathematics, and English) was electronically marked. One of the requirements of this project was the ability to compare online marking with paper-based marking, when the script had been marked by the same examiner. This means that five of the scripts that examiners marked on paper earlier were then directed to the same examiner via the e-marking system.

In his End of Session Report, Gary Black (2010) from RM Education noted that "despite what must have been a significant change for the examiners, the session ran incredibly smoothly" and that from the feedback received from examiners, the experience appeared to have been very positive for the majority of those who used the e-marking system.

In terms of marking quality, it was concluded that e-marking in all three subjects proved to be comparable to paper-based marking, and that examiners were marking to the standard set by the principal examiner (Slavec Gornik & Urank, 2009).

At the end of the project, the NEC and RM Education agreed that the pilot session ran very smoothly and that overall, the project aims and objectives were achieved. However, despite the positive outcomes of the project, the NEC did not obtain the agreement of the competent Ministry to implement e-marking to General Matura exams at that point.

# 3. Introduction of e-marking at the National Assessment in primary education

## The National Assessment in primary education: Slovenian context

In comparison with other countries of former Yugoslavia and some other former socialist countries, Slovenia has a relatively long tradition of external assessment. At the beginning of the nineties, we developed external assessment in primary education, and started working on Matura examinations in secondary education. The country's small size and finances were two main reasons why national assessment was organized as a centralized system from the start.

Since its introduction, external assessment in primary education has had different roles. From 1992 to 2005, it was a summative assessment, with a shared formative/summative function between 2002 and 2005, until it finally acquired a formative role from 2006 onwards. However, the results that ninth-graders achieve in L1 and mathematics can still be used, with parents' consent, as an additional criterion for admission to secondary schools with limited enrolment (Državna komisija za vodenje nacionalnega preverjanja znanja & Državni izpitni center, 2022).

The main goal of the existing National Assessment is to gain additional information on how well pupils at different stages of primary education attain the standards of knowledge set by the curricula. The National Assessment provides quantitative and qualitative data about pupils' achievements. Pupils can compare their individual achievements with those of their peers at school and with the national average. The National Assessment also enables teachers and schools to evaluate the quality of their work, while on the system level, it can be used as a basis for curriculum evaluation and for making further decisions about the development of the education system (Državna komisija za vodenje nacionalnega preverjanja znanja & Državni izpitni center, 2022).The costs of the National Assessment are covered by the national budget.

The National Assessment tests include different types of item formats, from MCQ, matching, sentence completion, short answers, etc. to open-ended extended writing tasks. The latter are less objective in terms of marking but provide invaluable information about specific competences that cannot be extracted in other ways.

## History of marking at the National Assessment

Before the introduction of e-marking, Slovenia had witnessed different organizational forms of external marking at the National Assessment, which was paper-based. At the time of 8-year-primary education, the marking was done by primary school teachers. Later, when the school system underwent a transformation, with primary education lasting 9 years, the NEC recruited external examiners for the marking of the National Assessment tests. In both cases**,** the examiners received payment for their work.

In 2006, when the National Assessment with a formative function was first administered, no longer having a decisive influence on final grades nor playing an important role for secondary school enrolment, the marking of national tests started to be considered as part of teachers' working duties. The Ministry considered that the formative function of the National Assessment should also apply to teachers: by going through the process of external marking, the teachers would learn about the importance of objective marking, and so the quality of internal marking would be enhanced as well. From then on, school principals appointed the teachers responsible for the marking of the National Assessment tests. With that, the marking stopped being compensated monetarily, which caused a lot of reluctance among teachers.

The marking of the tests at the end of grade 6 took place in schools; tests were marked by teachers, who used externally moderated marking schemes. Teachers also had to fill in marking sheets, which were then sent to the NEC where they were scanned.

The marking of the tests taken by ninth-graders took place at 17 marking centres around the country. The tests were marked by teachers using moderated marking schemes under the supervision of principal examiners' assistants. Marking sheets had to be filled in as well. Principal examiners' assistants were appointed by the NEC. The marking was managed by the National Education Institute (NEI), which also organized teacher training for the marking. The external marking of the tests in one subject was completed in one day.

The pupils and their parents were able to view the results and the marked tests. At the end of grade 6, this was done at schools in the presence of teachers who also carried out the re-marking procedure when enquiries upon results were made.

Enquiries upon results at the end of grade 9 were made by school principals at nine regional centres across the country, where they were resolved by principal examiners' assistants.

Organizationally, logistically, and security-wise, the whole marking process was very complex.

## Transition to e-marking at the National Assessment

Ever since the introduction of formative national assessment, marking has been one of the burning issues between teachers and school principals. The problem was largely discussed at the national level, by the competent Ministry and the NEC. Dissatisfaction with the existing marking system was also expressed by the NEI, which was responsible for the organizational side of the marking and for teacher training. There were several attempts to change the existing external marking system in order to abolish distinctions between the marking at the end of grades 9 and 6. At the NEC, we were aware that the marking should stay external, as the transfer of the marking to schools would lower the reliability and objectivity of test results. Intensive discussions about changes in the marking system started in June 2011. The NEC presented the concept of e-marking and the experience from the pilot project to the competent Ministry and to the NEI, and in September of the same year, all three institutions signed an agreement that in the school year 2012/2013, the National Assessment exams for ninth-graders would be marked electronically for the first time. The following year, e-marking would be introduced at the National Assessment for sixth-graders as well.

**The main reasons for the introduction of e-marking were thus as follows:**

### To prevent disruption in the school process.

School principals complained that the teaching process at schools was disrupted due to teachers' absence on the days when the National Assessment marking took place at the regional centres. Four days a year, the timetables had to be reorganized as the teachers of mathematics, Slovenian, and third subjects were absent, and replacements had to be found. School principals also complained about how enquiries upon results were carried out; they considered the procedure too expensive and time-consuming.

### To change teachers' attitude towards external marking.

The teachers found the existing marking system stressful and degrading. Marking the tests first and then transmitting the data on the scanning sheets afterwards seemed absurd to many of them. Besides, it often happened that they had to mark more tests than originally planned since some school principals did not send all available teachers to the marking centres. This caused feelings of discontent and of being unfairly treated among the teachers who were present at the marking centres and dutifully fulfilled their obligations.

**To provide as objective and valid results as possible.**

At the NEC, we wanted to improve and monitor the quality of marking to enhance the objectivity and validity of test results. The existing marking system allowed numerous technical errors, such as incorrectly filled marking sheets and miscalculations. For instance, the share of incorrectly filled marking sheets varied between 13% and 17%. Other problems included non-compliance with the marking scheme and difficulties in marking open-ended responses, which were a result of insufficient supervision and a lack of support offered to the examiners. Consequently, the number of enquiries upon results was high and was increasing every year. After remarking, the share of changes in results showed that originally marked scripts contained a lot of errors, which were not only technical in nature.

**To reduce the movement of scripts and increase security.**

Due to numerous transfers of scripts from schools to the NEC, from the NEC to the marking centres, and then back to the NEC and finally back to schools, there were problems with security. It happened almost every year that a script was lost, mostly during the marking process.

**To reduce costs.**

With the move to e-marking, the scripts would no longer need to be transferred so many times, and the shipment costs would decrease. E-marking would also remove the need for teachers and school principals to travel to regional marking centres, cutting their travel costs completely.

## Implementation of e-marking

### Preparation phase

After signing an agreement in September 2011 on introducing e-marking at the National Assessment in the following school year, the NEC considered two options: to develop its own e-marking software or to lease an established e-marking software. Certainly, our own e-marking software would offer more flexibility and independence in the long run. However, due to extreme time constraints, we opted for the lease. Functional software, compatible with our needs, seemed like a much more reliable option at the time. In March 2012, a tender was published, and the company RM Education with its Scoris Assessor (now RM Assessor) application was selected as the provider. Having successfully cooperated with RM Education before (pilot e-marking project in 2009), the choice of provider was a logical one. All formal procedures were finished by August 2012, when intensive preparations for the introduction of e-marking began, as seen from the table below.

| Activity | Date |
|---|---|
| Preparation of Project Initiation Document | August 2012 |
| Preparation of timeline | August 2012 |
| Preparation of e-marking software | August–November 2012 |
| Installation and configuration of software | August–September 2012 |
| Translation of user interface | August–September 2012 |
| Translation of user guides (marking user guide, standardisation setup guide, supervision guide, online training) | September 2012–January 2013 |
| Preparation of new software and adjustments to NEC information system (development of interface, indexing software, teacher registration software, enquiry upon results software) | October–December 2012 |
| Test preparation (setting up new formats, page codes, defining test structures in Scoris Assessor) | September 2012–March 2013 |
| Preparation of digitalization procedures | October 2012–March 2013 |
| Pilot project (test preparation, trialling and confirming the pilot version of the software, pilot standardization and marking) | October 2012–December 2012 |
| Informing schools, pupils, parents, and the public about e-marking | September 2012–June 2013 |
| Preparation of instructions for schools on how to carry out e-marking and enquiries upon results | September 2012–March 2013 |
| Workshops and trainings for NEC employees, computer operators, e-marking assistants at schools, NEI advisors, principal examiners, and principal examiners' assistants | September 2012–April 2013 |
| Defining procedures for teacher/examiner data recording; database setup and database management | September 2012–January 2013 |
| Defining conditions for appointing principal examiners' assistants, deciding on selection procedure and the number of assistants needed, arranging database | September 2012–March 2013 |
| Preparation of timeline for e-marking and enquiries upon results | February–March 2013 |
| Preparation of online questionnaires for school principals, teachers/examiners, principal examiners, and principal examiners' assistants | March–April 2013 |
| E-marking (May session)<br>E-marking (June session) | 13 May–24 May 2013<br>3 June–6 June 2013 |
| Enquiries upon results (May session)<br>Enquiries upon results (June session) | 29–31 May 2013<br>11–12 June 2013 |
| Analyses of questionnaires | June–August 2013 |
| Preparation of reports and analyses | May–September 2013 |

Table 1: Sequence of activities related to the introduction of e-marking

From signing the license agreement to establishing the e-marking system, it took only eight months. The NEC was aware of the fact that the project would only be successful if there were no major technical issues, if the training network was well organized and if the users' experience with the application was mostly positive.

The data from previous years showed that there were around 4,000 teachers who would be involved in e-marking. RM Education delivered initial trainings on the use of the e-marking application and on the standardization process for NEC employees, principal examiners, and computer operators responsible for teacher trainings. RM Education also carried out workshops on system administration, the use of the administrative interface, technical support, and online services. The NEC provided additional workshops on the use of the Slovenian version of Scoris Assessor for computer operators and prepared them for further trainings that they had to carry out for 463 e-marking assistants appointed by school principals at schools across Slovenia. The role of e-marking assistants was to ensure a smooth e-marking process at schools and to familiarize teachers with the technical side of the e-marking application. For that purpose, we also had to set up a Slovenian version of the Familiarisation Mode of the application, which was used by all parties in the training network. It enabled the users to familiarize themselves with the application and to try out different tools that the application offered.

In November and December 2012, the NEC and subject testing committees tried out the pilot Slovenian version of the application. We tested the basic platform for examiners. However, due to time constraints, it was impossible to test all features available (script review, enquiries upon results, marking of atypical scripts, etc.)

The NEC also provided training for 34 subject advisers at the NEI, who later trained more than 4,000 teachers at 258 workshops for the practical use of the application. The workshops were organized in small groups so that each participant could use their own computer. In March and April, the NEC organized trainings for 225 principal examiners' assistants, who participated in the pilot standardization setup process together with subject testing committees.

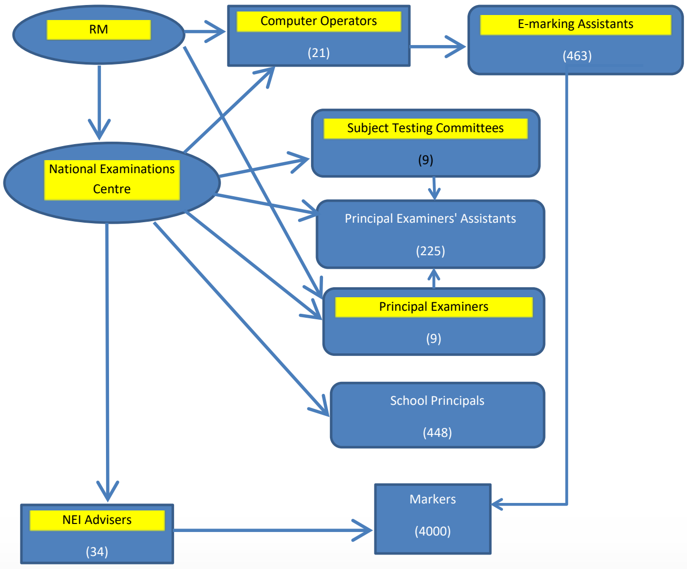


Figure 1: Organisational outline of training for the e-marking process

### Implementation phase

In May 2013, after ninth-graders had taken the exams on paper, their scripts were sent to the NEC where they were scanned and uploaded as digital images to the e-marking system. Moderation of marking schemes was done by expert teams, which consisted of subject testing committees' members (including the principal examiner) and principal examiners' assistants. In addition, a so-called standardization setup in Scoris Assessor application had to be prepared. Expert teams had to choose one practice, two standardization and three seeding responses and mark them, thus applying so-called definitive marks, which would later serve as a quality control mechanism during online marking. The NEC was given 4 to 6 days for the above activities, which often overlapped. We also had to prepare the final marking hierarchy in our database, which would then be transferred into the e-marking application. Each subject would have the same hierarchy, consisting of a principal examiner, the principal examiner's assistants/team leaders (around 50 for major subjects such as mathematics or Slovenian) and examiners (around 1000 for major subjects). For major subjects such as mathematics or Slovenian, each team leader would be responsible for a group of around 20 examiners. The only exception were subjects with a very small number of candidates (Italian, Hungarian, and Lower Education Standard), which did not have principal examiner's assistants, and so the examiners were directly monitored by the principal examiner.

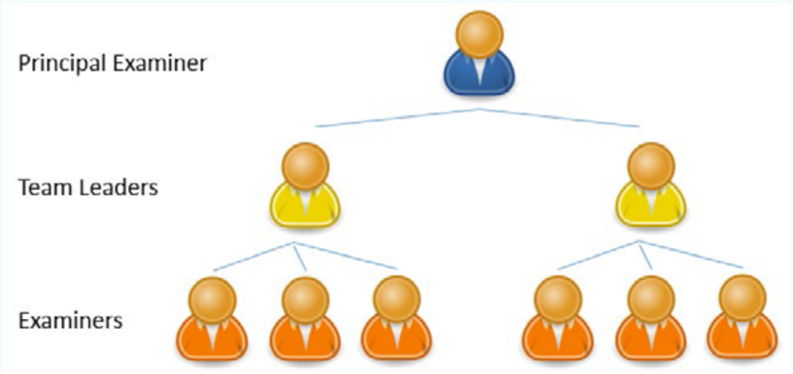


Figure 2: Hierarchical structure of examiner team (RM Education, 2022)

After all scripts were digitalized and the e-marking system was filled with the necessary data, the e-marking process could begin. Teachers of mathematics and L1 were given five working days for e-marking, while three working days were assigned for e-marking of other subjects.

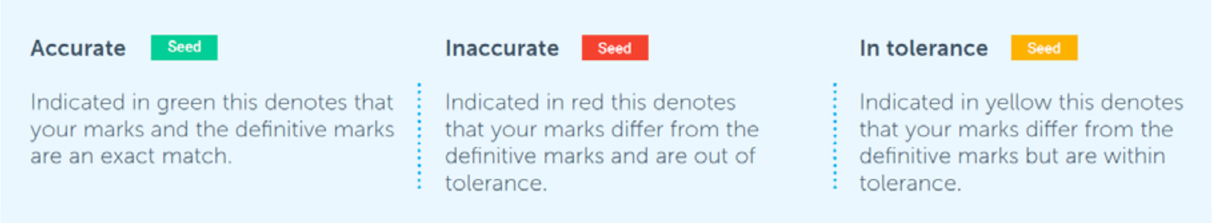
Principal examiners in each subject and their assistants were responsible for providing expertize, monitoring examiners' work, and ensuring accuracy. There were around 50 principal examiner's assistants for mathematics and Slovenian, 45 for English, and 25 for each of the following subjects: geography, history, design and technology. As for the number of examiners and the number of scripts that each examiner had to electronically mark, the data from the NEC database shows the following:

| Subject | No. of candidates | No. of examiners | No. of scripts per examiner |
|---|---|---|---|
| mathematics | 17,760 | 1,012 | 18 |
| Slovenian (L1) | 17,695 | 1,036 | 17 |
| Italian (L1) | 50 | 7 | 7 |
| Hungarian (L1) | 15 | 5 | 3 |
| English (L2) | 4,370 | 836 | 5 |
| German (L2) | 253 | 38 | 7 |
| geography | 4,309 | 354 | 12 |
| history | 4,627 | 359 | 13 |
| design and technology | 4,201 | 361 | 12 |

Table 2: No. of candidates, examiners and scripts per examiner at different subjects at the National Assessment in grade 9, 2013 (NEC database, 2013)

The first day of marking was devoted to practice and standardization. By marking a so-called practice response examiners familiarized themselves with the marking scheme and marking instructions. Examiners were then required to mark two standardization responses before moving on to so-called live marking, when they could download, mark, and submit other responses. If the standardization responses were not marked within the required tolerance, team leaders provided feedback to the examiners; they pointed out marking errors and deviations from definitive marks. Within the e-marking application, examiners themselves were also able to compare their marks with definitive marks.

After submitting the standardization responses, all examiners were approved for marking. Standardization process was not used to suspend examiners or make them undergo a second standardization. It only served as a soft quality control mechanism. If we had used standardization as a suspension tool, we would have risked that teachers, who were not paid extra for marking, would poorly mark the standardization scripts deliberately in order to get suspended. Nonetheless, we did use seeding responses (seeds) to keep marking quality consistent during live marking. Seeds are definitively marked scripts that are randomly and anonymously added to examiners during live marking. Team leaders, as well as the examiners themselves, are able to see how accurately the seeds were marked.



Accurate 🟩 Seed

Indicated in green this denotes that your marks and the definitive marks are an exact match.

Inaccurate 🟥 Seed

Indicated in red this denotes that your marks differ from the definitive marks and are out of tolerance.

In tolerance 🟨 Seed

Indicated in yellow this denotes that your marks differ from the definitive marks but are within tolerance.

Figure 3: Accuracy indicators for Seeds (RM Education, 2022)

After e-marking was finished, ninth grade students were able to electronically view their marked scripts. For the first time in the history of external assessment in Slovenia, students and their parents had online access to scripts by using their personal identification number and their exam code. If the students, their parents, or their teachers discovered errors in marking, or thought that certain responses should be remarked, school principals used a special application, where they filled out an online form to appeal against the results. The principal examiners and their assistants then had two days to review the enquiries and re-mark the responses in question. After that, the final results were published.

### Difficulties during preparation and implementation phases

Due to the professionalism of the RM Education and NEC teams that worked on e-marking in the first year of implementation, we did not encounter any major problems. Considering the very short deadlines that we had to meet, we were happy to note that the preparation and implementation phases had been carried out extremely well. Nevertheless, there were certain difficulties that we had to deal with:

- During the preparation phase, it was not possible to test all procedures and tools in the e-marking application. This means that they were not tested until they went live. Therefore, there was an ongoing feeling of uncertainty, but fortunately, it all ran smoothly in the end.
- Test papers were prepared before the decision about the implementation of e-marking was taken; thus, they were only partly adapted to the new requirements. If item-writing teams had been able to use the e-marking application beforehand, they would have probably made some additional changes to the tasks, such as different layout, more spacing between test items, etc.
- Some examiners were not well prepared for e-marking. This was due to their absence from the trainings and to the fact that some teachers lacked basic ICT competencies.
- Some examiners, who were not marking to the standard, ignored the messages sent by their team leaders. This left team leaders with feelings of frustration, as they knew that some students would receive inaccurately marked scripts.
- The marking of unscannable scripts (A3 enlargements, braille, latex, etc.) represented a special challenge. Such scripts were marked in written form by the principal examiners' assistants; later, the scores were entered into the e-marking application.
- There was some anger on the part of some examiners about the seeding scripts; they did not understand the purpose of this quality control mechanism although it was explained in the user guide.
- In several open letters, addressed to the competent Ministry, general public and to everyone involved in the e-marking implementation, some teachers complained about spending money on e-marking at the time of financial crisis. As this was happening right before we were about to go live with e-marking, it was very stressful for everyone involved.

# 4. Positive outcomes of implementing e-marking at the National Assessment

For the NEC, improving marking quality was one of the principal reasons for introducing e-marking. By no means did we expect to see changes in the first year, when our primary task was to ensure that the system worked smoothly. However, certain progress was already noted after the very first e-marking session was finished.

Over the years, after implementing e-marking for exams at the end of grade 6 as well, and having used the RM e-marking application for multiple exam sessions, we can highlight four main advantages of e-marking:

### 1. Better marking quality

The e-marking application does not require examiners to manually record marks as the paper-based process did, which means that scanning sheets are no longer part of the process. Technical errors such as miscalculations are thus eliminated as the e-marking application automatically calculates the total score. Enquiries upon results based on calculation errors are now a thing of the past.

### 2. Improved examiner monitoring and feedback during the e-marking process

Automated assessment of standardization and seed marking, combined with ongoing monitoring done by team leaders and principal examiners, offers a deep insight into marking quality. Examiners are thus offered improved real-time feedback. During the entire e-marking process, they are able to communicate with their team leaders via an in-app messaging system. They can also escalate so-called exceptions, which are then swiftly resolved. Examiners receive clarifications about marking dilemmas from principal examiners or their assistants. It is possible to detect and resolve marking inconsistencies as they happen, rather than as the result of enquiries afterwards. Consequently, we have seen a lower share of changes in marks following enquiries upon results in the years following the implementation of e-marking.

### 3. Possibility of advanced analyses

The data, which is collected after an e-marking session is finished, allows researchers at the NEC to analyze marking quality at a granular level. Reports based on this data are then sent to school principals, who analyze them together with examiners.

Besides, digitalized student responses are incorporated into the program called OrKa, which was developed at the NEC. It is a tool designed for school professionals to view students' achievements and to access the National Assessment data at the school and national level. The data is used to capture granular information about student performance as the tool offers an insight on how the students performed on the level of individual test items.

Teachers can use OrKa to prepare task- and item-based analyses and review individual test items along with digitalized student responses. This process helps them identify strong and weak areas in students' knowledge.

### 4. Positive feedback from school principals and teachers

For several years following the implementation of e-marking, school principals and teachers were asked to complete a questionnaire prepared by the NEC after every exam session. The questionnaire contained questions about e-marking trainings and e-marking itself, about communication with team leaders or with the principal examiner, about quality control during e-marking, and about the advantages and disadvantages of e-marking.

A crucial question in the first year of implementation was: What was your opinion on e-marking before it was implemented, and what is your opinion after having participated in the e-marking process?

The analysis, based on 412 responses by school principals (98.8%), showed that more than half of school principals were in favour of e-marking before it was introduced, with support rising to 82% after the implementation.

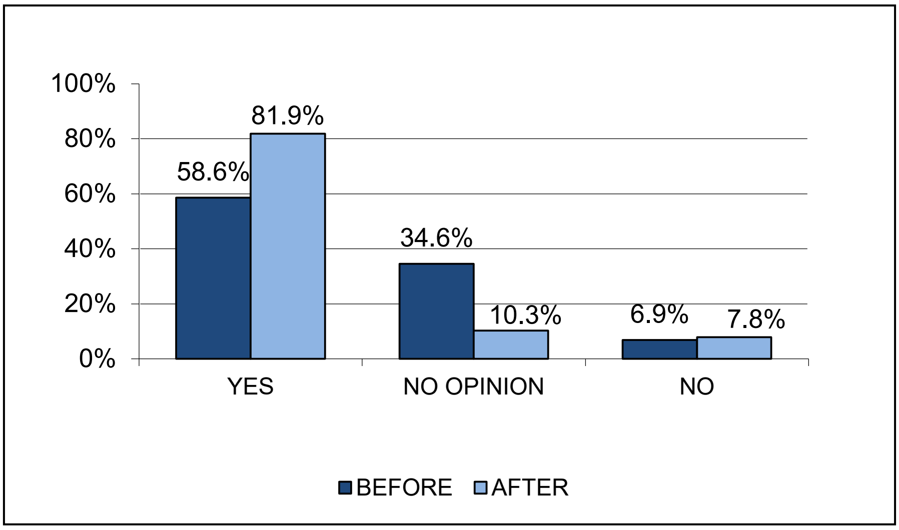


Figure 4: School principals' support for e-marking before and after the implementation (Semen, 2013)

A teacher survey was completed by 1,567 examiners, which represents 40% of all examiners. Despite the fact that teachers expressed less support compared to school principals, a big difference in opinion before and after the introduction of e-marking could be seen. Teachers' support before the implementation of e-marking was low, amounting only to 21%. After the implementation, it rose to 56%.

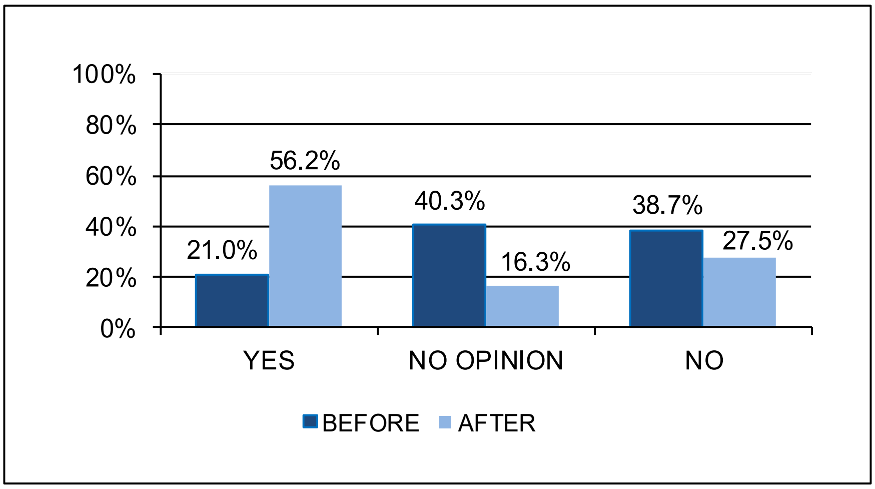


Figure 5: Teachers' support for e-marking before and after the implementation (Semen, 2013)

In the year 2017, after school principals and teachers had already participated in the e-marking process in four subsequent years (and after teachers started receiving extra payment for e-marking in 2015), the questionnaire analysis showed that a vast majority of school principals were in favour of e-marking. As the main advantages, they stated greater flexibility in time management and electronic access to scripts. They also appreciated the fact that e-marking did not disrupt the normal school process. The teachers gave very high marks to functionalities within the e-marking application. Most of them expressed satisfaction with the fact that their marking was monitored by team leaders and were content with the effective and polite communication that they experienced via the e-marking messaging system. The leading advantage that the teachers saw in e-marking was the freedom to choose when and where to mark. The fact that they no longer had to calculate final scores and that the system warned them if they left any responses unmarked also ranked highly in their list of advantages. As a main disadvantage, however, a lack of personal contact with other examiners was stated (Semen, 2017).

## 5. Gradual transition to e-marking at the General Matura

Having successfully implemented e-marking on the primary level, the goal of the NEC was to do similarly on the secondary level, at least for the General Matura. The Vocational Matura was not included in the original plans since the marking of the Vocational Matura exams is organized differently.

The project of gradually introducing e-marking to the General Matura was financed by European Cohesion Funds. In the years 2016 to 2022, relying on the experience of implementing e-marking at the National Assessment, we successfully managed to implement e-marking for all General Matura exams.

The first six exams were electronically marked in 2017, and by 2021, about six new exams were added every year. From an organizational point of view, the main difference between the National Assessment and the General Matura lies in the number of different exams per session; at the National Assessment, this number is 21, while at the General Matura the quantity of exams amounts to 47. However, since there are fewer candidates at the General Matura than at the National Assessment, the total number of scripts is lower (about 77,000 compared to around 100,000 at the National Assessment). Consequently, the number of examiners is lower as well (around 1,000 examiners at the General Matura compared to around 7,000 examiners at the National Assessment).

Due to the significantly higher number of different exams, the adaptation of test papers for e-marking at the General Matura was more time-consuming. Compared to the National Assessment, which is a low-stakes examination, the General Matura is a high-stakes examination. That is why more complex marking procedures had to be introduced, such as double standardisation, control marking consisting of double and third markings, candidates-at-risk marking, and other types of control marking. Test items at the General Matura are also more difficult to mark compared to those at the National Assessment, as there is a larger proportion of structured items and essay questions.

Despite all the challenges mentioned above, e-marking was successfully introduced for all written exams at the General Matura. Surveys and analyses related to e-marking have confirmed that users in secondary education (i.e., teachers and external examiners) have embraced the online environment as a tool that efficiently provides better support for their work, similarly to their colleagues in primary schools.

The main objectives of the project were to enhance the quality, objectivity, and reliability of marking and to increase security during the marking process. Since external examiners are mostly secondary school teachers, their use of the e-marking platform also contributed to improving their ICT competencies. Additionally, the e-marking system was further upgraded by features, internally developed at the NEC, such as a platform for e-appeals. The successful achievement of the above goals was closely connected with the development and implementation of various procedural, administrative, and technical solutions. As part of the project, we also conducted an analysis of the feasibility of introducing e-marking for Vocational Matura exams (Državni izpitni center, 2022).

## 6. Conclusion: challenges for the future

When discussing long-term objectives for the future, it is easy to relate to the views expressed in the AQA publication (2015) The Future of Assessment 2025 and Beyond: "It's about acknowledging that our imperfect system has served us well for 30 years – but it will, at some point, need to change if it is to continue to support our young people's education…"

For external assessment in Slovenia, moving to e-marking was probably just a first step. A transition to electronic examinations, where students would no longer hand-write their exam responses on paper, could be a logical next step. In theory, at least.

In practice, a transition to e-testing is extremely demanding, which we experienced in the post-pandemic year of 2021 when NEC carried out a pilot e-testing project together with RM Education.

The project included all Slovenian primary schools; around 47,000 students in grades 6 and 9 participated in taking a very basic e-test, consisting of several mathematics, L1 and L2 items. After the project was finished, a number of doubts emerged because we realized the following facts:

- Primary schools in Slovenia are not equipped with an adequate number of computers that would allow us to carry out a large-scale assessment for all schools simultaneously.
- There is heterogeneity in hardware and software infrastructure.
- Many schools still have problems with internet connection.
- Students' and teachers' ICT competencies should be enhanced in order to reach the level required for competently participating in e-assessment.
- There is a lot of scepticism concerning e-testing among teachers.
- Until e-testing is not part of regular school practice, it cannot be implemented on a national level.
- Developing online tests is not about replicating paper exam; it requires technical know-how and specific skills that our test developers lack.
- Online test items can include complex interactive elements, animations, audio and video clips; it is crucial that item writers change their traditional perspective on developing test items and that they have an understanding of what they want to test with a particular item.
- There are many security issues that would need to be tackled.

(Državni izpitni center, 2021)

All of the above has left us apprehensive about the transition to e-testing. At the same time, we do agree with AQA's Chief Executive Andrew Hall (2015) that technology offers the opportunity to take the validity and reliability of assessment to new heights.

At the moment, the more realistic short-term challenges for the NEC are as follows:

- To implement e-marking at the National Assessment in grade 3, which will become compulsory in the school year 2024/25.
- To strive for a gradual transition to e-marking at the Vocational Matura.
- To enrich our existing reporting tool for schools, which supports an evidence-based approach, used for improving the quality of teaching and learning.

## References

AQA (2015). *The Future of Assessment 2025 and Beyond (online).* Available: https://filestore.aqa.org.uk/content/about-us/AQA-THE-FUTURE-OF-ASSESSMENT.PDF

Black, G. (2010). *RIC (Slovenia) E-Marking Pilot Project, November 2009. End of Session Report.* (Project report, RM Education)

Državna komisija za vodenje nacionalnega preverjanja znanja, Državni izpitni center (2022). *Izhodišča nacionalnega preverjanja znanja v osnovni* šoli *(online)*. Available: https://www.ric.si/nacionalno-preverjanje-znanja/splosne-informacije/

Državni izpitni center (2022). *Zaključno poročilo projekta »Priprava in implementacija e-ocenjevanja pri maturi*«. (Project report, National Examinations Centre)

Državni izpitni center (2021). *Poročilo o izvedbi pilotnega projekta »E-testiranje v osnovni* šoli«. (Project report, National Examinations Centre)

RM Education (2022). *RM Assessor 3 Supervision Guide: V3.2*

RM Education (2022). *Marking User Guide V4.8*

Slavec Gornik, A., Urank M. (2009). *Pilotni projekt elektronskega ocenjevanja pri splošni maturi.* (Project report, National Examinations Centre)

Semen, E. (2017). *Analysis of questionnaire on e-marking at the National Assessment for school principals*. (National Examinations Centre)

Semen, E. (2017). *Analysis of questionnaire on e-marking at the National Assessment for teachers.* (National Examinations Centre)

Semen, E. (2013). *Analysis of questionnaire on e-marking at the National Assessment for school principals*. (National Examinations Centre)

Semen, E. (2013). *Analysis of questionnaire on e-marking at the National Assessment for teachers*. (National Examinations Centre)

# Sweden

# SWEDEN: BIOGRAPHIES

## Karin Hector-Stahre

Karin Hector-Stahre is a Head of unit at The Department for National Curricula at the Swedish National Agency for Education. Her main focus is and has been assessment including national tests, the development of national digital tests and most recently the preparation for the introduction of centrally assessed national tests. Karin holds a teacher-certificate in Swedish and English for upper secondary school.

# THE EXPERIENCE OF DEVELOPING AND LAUNCHING E-TESTING ON A NATIONAL LEVEL

## Abstract

Sweden started introducing digital national tests on a national level in the spring of 2024. In this paper, the process of developing and gradually introducing e-testing is described.

The article describes several challenges in starting up a system for digital national tests available for all students within the last year of compulsory school, grade 9, upper secondary school, and adult education at the upper secondary level.

Particular attention is given to the adaptation of the user interface in the digital assessment platform. The interface has been comprehensively adapted to comply with the legislation for accessibility.

The task to increase the preparedness of schools and school organisers in implementing the use of digital national tests is described as well as the challenges to implementing such tests in a decentralised school system.

Another issue addressed in the article is the gradual development of external scoring and assessment, which will entail the recruitment of over 3,000 certified teachers.

## What we learned from developing and launching e-testing on a national level

In 2017, The Swedish National Agency for Education (Skolverket)[1] received a government assignment to develop digital national tests.[2] As a result of this task, Sweden will be introducing digital tests nationwide during the spring of 2024. This article focuses on and describes the main challenges of the assignment during the period 2017 to 2023. When this yearbook is published, the entire system has recently been launched and there certainly will remain questions to be answered.

Particular attention will be paid to the adaptation of the user interface in the digital assessment platform and the preparedness of schools and school organisers.

[1] Skolverket is the central administrative governmental body for the public school system, publicly organized preschool, school-age childcare and for adult education.

[2] This is the Swedish National Agency for Education (Skolverket.se).

## Background

Even though Sweden has a decentralized education system, schools and school organisers must adhere to goals and learning outcomes that are defined at a central level. The government has the overall responsibility and sets the framework for education at all levels.[3]

Municipalities and independent school providers are the principal organisers in the school system; they allocate resources with additional funding from the Swedish state and organize activities to ensure that the students can reach the national goals.

Based on this, each school chooses the working methods most appropriate for them. The work is followed up using a systematic quality assessment conducted by the Swedish Schools Inspectorate. In its supervision, the Inspectorate checks that the activities under review meet the requirements set out in laws and other regulations. Supervision results in assessments of whether there is a deficiency concerning the requirements of the regulations. The Schools Inspectorate makes decisions directed at the school authority to take the measures necessary to meet the requirements. In quality review, the quality of the education or activity is assessed with goals and other guidelines. The School Inspectorate assesses the school's or principal's work against the quality criteria set by the authority and expresses its opinion on the quality level. The reports are available to the public.[4]

## Number of schools

There are 290 municipalities in Sweden. In the school year 2022/2023, there were 4,719 compulsory schools in total, of which 3,868 were municipal schools and 831 were independent schools. In the same school year, there were 1,295 upper secondary schools of which 817 were municipal upper secondary schools and 466 were independent upper secondary schools.[5]

### Independent schools

School reforms in the 1990s transferred the power to choose school from the public authorities to the individual family. The school market that emerged also created differences that challenge equivalence in Swedish schools. The independent schools are financed by taxes and there are no fees. The term independent education provider refers to an independent physical or legal entity that operates education at an independent preschool or school. This could be, for example, an association, foundation, or limited company. [6]

[3] (Eurydice, 2024). Overview (europa.eu).

[4] English (Engelska) (skolinspektionen.se).

[5] siris.skolverket.se.xls (live.com).

[6] Skolverket 2014.

## Exams, standardized tests and national tests

The forms and aims of central examination and testing have gradually changed from the 1960s and onwards due to education reforms and new curricula. Until 1968, Sweden had so-called matriculation exams at the end of upper secondary school. These exams were conducted by external examiners and teachers, and the result of the exams decided whether or not a student would receive a diploma.

So-called standardized tests in compulsory schools were introduced in the 1940s and served as support for teachers in their assessment of students as well as in the implementation of new curricula. They were conducted from grade 2 to grade 9 and included Reading, Writing, Mathematics, English, German, and French. This kind of standardized test continued until 1983 (Lundahl, 2009).

In upper secondary school, central external examinations were replaced by central tests in Swedish, English, Mathematics, Physics and Business Administration, German, French, and Accounting.[7] These tests were constructed at Skolöverstyrelsen, the national agency that preceded the current Skolverket. The role of these tests was to support teachers in the grading process but also to regulate grading in the test subject and in other subjects as well. The tests were assessed by the pupils' own teacher.

The introduction of a new curriculum in 1994 entailed two changes in the school system in connection to assessment: new principles for grading and a new form of national tests in Swedish, Swedish as a second language, English and Mathematics. These new national tests were conducted in both compulsory school and upper secondary school. The tests had several aims e.g. to support grading and assessment, demonstrate strengths and development needs, concretise the curricula, and could also be used as a basis for analyses of whether goals are met at school both on a local and national level. While earlier tests had regulated grading to some extent, some of these new national tests were optional and were meant to support the grading process rather than control it. Since the abandonment of the matriculation exams in the 1960s the function of national exams has been to support teachers' grading. No singular test och exam has since had the role of selection when transitioning to the next level of education.

In the last 30 years, the number of national tests has increased, particularly in compulsory schools with the introduction of national tests in grades 3 and 6, and new test subjects.[8] Since 2018, the results of national tests must be given extra consideration before teachers grade a course or subject. In other words, the test result should play an important role when teachers grade a course or a subject, and it now has a greater significance than other bases for assessment. However, the national test should not entirely determine the course or subject grade, and the result of a national test cannot be the teacher's only basis for grading. There is no set relationship regarding what impact the grade from the national test should have on the final course or subject grade. It is however less likely that a pupil is awarded a very high final course or subject grade if the grade on the national test was very low and vice versa. The principal is responsible for monitoring the grading concerning the test results, but the individual

[7] Ibid.

[8] History, religion, geography, civics, chemistry, physics, and biology in both grade 6 and grade 9.

teacher gives the final grade based on the collective result for each pupil.[9] The fact that it is the teacher alone who is responsible for the final grading is seen as one of the challenges to the national equivalence in grading. One measure to deal with this is the introduction of a central assessment. There is also a governmental investigation that should give proposals for changes in the grading system. The proposals are intended to ensure that grades from elementary school and upper secondary level more fairly reflect students' subject knowledge and to counteract grade inflation.[10]

## Introducing digital national tests

Digitization has generally come a long way in Swedish schools, but it does not look the same across the country.[11] The starting point for e-assessment at a national level is a government assignment given to Skolverket in 2017 (Skolverket 2020).[12] In the same year, the Government decided on a national digitization strategy for the school system, which states that the aim of the digitization policy is that Sweden become the best country in the world when it comes to taking advantage of the possibilities that digitization presents. Before this assignment, there were no comparable initiatives at a national level. within the assessment field. It is worth noting that e-assessment is often used locally when schools develop their own tests.

## The organisation of national tests in Sweden

National tests are conducted in compulsory school, upper secondary school and within adult education.

| Grade 3 | Grade 6 | Grade 9 | Upper Secondary | Adult Education |
|---|---|---|---|---|
| Swedish | Swedish | Swedish | Swedish (2 courses) | Swedish (2 courses) |
| Mathematics | Mathematics | Mathematics | Mathematics (4 courses) | Mathematics (4 courses) |
| | English | English | English (2 courses) | English (2 courses) |
| | | Science (biology, physics, chemistry) | | |
| | | Social sciences (religion, history, geography) | | |

[9] Genomföra och bedöma prov i grundskolan - Skolverket.

[10] Likvärdiga betyg och meritvärden - Regeringen.se.

[11] Uppföljning av digitaliseringsstrategin.

[12] Uppdrag att digitalisera de nationella proven m.m. - Regeringen.se.

Beginning in grade 6, students receive grades once a semester within a two-semester system. In grading, the teachers assess what knowledge the pupil has demonstrated during the semester. Final grades are given when the pupil has completed the study of all the subjects included in compulsory school. This takes place when grade 9 is concluded. Schools also employ formative assessment to provide continuous feedback to students throughout the school year.

In upper secondary school and adult education, grades are given after each completed course. In 2025 the course system will be replaced by a new system where all curricula are designed as cohesive subjects with levels instead of courses. A subject can have one or more levels. In many subjects, students will be able to receive instruction for a longer period before the final grade is set. This means that the students should be given more time to immerse themselves in a subject before the final grade is set. The teacher is also given better opportunities to conduct teaching in the long term and based on a holistic approach.

The grade will better reflect what the student knows at the end of their studies in the subject.[13]

Students' responses are assessed by the students' own teachers and the results will only provide support for the teachers' grading. Hence, the national tests do not play the role of exams in the traditional sense. However, so far there is no formal regulation that stipulates to what degree the result of a national test should affect the final grade. The student participation in the national tests is mandatory. Since the tests are compulsory and result of the test is an important part of the course or subject grade a majority of the students are motivated to participate. In some cases, however, there may still be reasons to exempt a student from a national test. So-called special reasons are required to exempt a pupil and it is the head teacher who decides whether a pupil should be exempted from taking a national test in whole or in part. Special reasons may be, for example, that the student lacks the knowledge of the Swedish language required to be able to take a test.

## The government assignment

The task given to Skolverket by the government stipulates that digital national tests are to be available in both compulsory school and upper secondary education.[14] [15] It also states that Skolverket should strive to increase the number of items that can be assessed automatically.

Rationales for the assignment are:

- Equity in assessment
- Fair grading
- Reduce teachers' workload.

[13] Skolverket 2024. Aktuell information om Gy25 - Skolverket.

[14] Uppdrag att digitalisera de nationella proven m.m. - Regeringen.se.

[15] National tests for grade 3 in compulsory school are excluded and will still be given on paper.

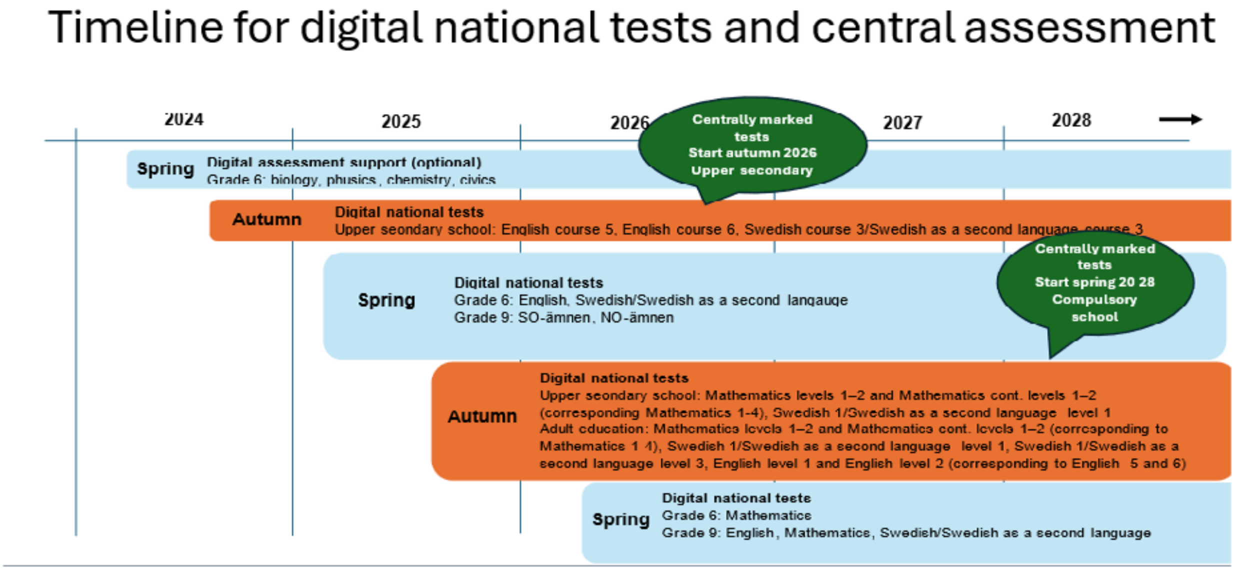
The assignment includes digitisation of both national tests and on-demand tests. According to the plan, Skolverket will introduce assessment support in the form of on-demand digital tests in the spring semester of 2024, and digital national tests for upper secondary school in the autumn semester of 2024.

Possible positive outcomes of introducing digital national tests and assessment support:

- A large proportion of automatically corrected student responses and anonymised student performance leads to more equitable assessment.
- Increased usability and accessibility for pupils with disabilities can be attained through technical solutions.
- Data management becomes more efficient.
- Paperless distribution contributes to a more efficient, safer, and more sustainable handling of the tests.

## Tests in the platform

The introduction of digital national tests will be done gradually. The first tests that will be conducted within Skolverket's digital test system will be English, Swedish, and Swedish as a second language in upper secondary schools in the autumn of 2024. In the spring of 2025, digital national tests will be introduced in grade 9 in English, Swedish, Science and Social sciences.[16] For those schools that either are not prepared from a technical perspective, or that experience technical issues in connection with digital national tests, Skolverket will be providing replacement tests on paper for a period of two years.



Timeline for digital national tests and central assessment

2024
2025
2026
2027
2028
Centrally marked tests Start autumn 2026 Upper secondary
Centrally marked tests Start spring 20 28 Compulsory school
Spring
Autumn
Spring
Autumn
Spring

Digital assessment support (optional)
Grade 6: biology, phusics, chemistry, civics

Digital national tests
Upper seondary school: English course 5, English course 6, Swedish course 3/Swedish as a second language course 3

Digital national tests
Grade 6: English, Swedish/Swedish as a second langauge
Grade 9: SO-ämnen, NO-ämnen

Digital national tests
Upper seondary school: Mathematics levels 1–2 and Mathematics cont. levels 1–2 (corresponding Mathematics 1-4), Swedish 1/Swedish as a second language level 1
Adult education: Mathematics levels 1–2 and Mathematics cont. levels 1–2 (corresponding to Mathematics 1 4), Swedish 1/Swedish as a second language level 1, Swedish 1/Swedish as a second language level 3, English level 1 and English level 2 (corresponding to English 5 and 6)

Digital national tests
Grade 6: Mathematics
Grade 9: English, Mathematics, Swedish/Swedish as a second language

[16] Biology, chemistry, and physics respectively civics, history, geography and religion.

The students will complete the digital national tests in a new test platform. When the students have completed the tests, some test assignments will automatically be corrected directly in the test platform. The test assignments to be assessed by teachers are also handled directly in the test platform. Once the assessment is complete, the test results will be available to the schools on the platform itself.

### Automatically corrected assignments and assessment

The goal is for the national tests to contain as large a proportion of automatically corrected test items as possible, i.e., the student answers are automatically corrected in the test platform. Parts of the tests that are automatically corrected are, for example, items with short constructed responses and multiple-choice items where the student chooses one of several alternatives. These types of items are already included in the national tests that are carried out on paper. When the tests become digital, the proportion of this type of response will increase. Automatically corrected responses help to both streamline the assessment work and increase the reliability of the assessments.

Tasks where students are asked to create their own responses or solutions will be assessed manually in the test platform, with the exception of some response items. Assignments where students are asked to reason or write a longer text or essay in which they can argue a point or the like will be assessed manually by teachers directly in the test platform. In other words, the rater will access the texts in the test platform and assess them there.

The number of tasks and the time dedicated to each test varies between subjects, courses, and school year. Within the present system with paper-based tests, the time for the written parts can vary from 60 to 120 minutes. Initially, there are no plans to change the time scope because of the introduction of e-testing. Compared to a few years ago the number of open-ended tasks is decreasing while the longer open-ended tasks and essays are still an important part of the tests as they reflect the subject syllabi. To support the teachers in assessing the test the universities develop test-specific rubrics and instructions for correcting and assessing the tasks.

## Pilot schools

Within the scope of the government assignment, Skolverket had the opportunity to work with 100 test schools.

The provision of test schools was secured by a government regulation that stipulated that 100 schools were to be appointed by Skolverket to function as pilot schools for the implementation and introduction of digital national tests. The selection of schools was done in such a way as to ensure variation between municipal and independent schools, and between larger and smaller schools.[17]

[17] Approximately 100 school units were included, and Skolverket made an appropriate selection of school units based on factors such as school type, year and type of school organiser. The selection includes the following types of school: compulsory school, special school, Sami school, upper secondary school, and municipal adult education at upper secondary level.

The agency was able to work with the schools from the start of the government mission, i.e. 2017. The different interactions with the schools have played an important part in Skolverket's work with the mission, although it became evident that the pace of the development of the test service was difficult to synchronize with its use in test schools. Delays in the development of the test service led to in-depth technical tests of the system being carried out towards the end of the test-school regulation, which ended on 31 December 2023.

# Challenges

## Accessibility and usability

The test service, including the test platform, must meet the legal requirements for accessibility to digital public services. Together with the supplier of the test platform, Skolverket has prioritized compliance in this area.

In the transformation from national tests on paper to digital national tests, there was an opportunity to focus on accessibility and usability.[18] This was valid both for the technical and content development of the tests. The concept of user-centred design was applied in the development of the test service, with the aim of benefiting students as well as other users of the system such as teachers and administrators.

Accessibility compliance posed a challenge for the universities that were tasked by Skolverket to design digital national tests.[19] They were faced with the challenge of converting paper-based concepts to new rules, regulations, and templates tied to the digital transformation of the national tests.

Skolverket has developed its own student interface to ensure that the test platform meets the requirements for accessibility. Other interfaces in the platform cannot be adapted specifically for Skolverket. On the other hand, the Web Content Accessibility Guidelines (WCAG 2.1 AA)[20] also apply to these interfaces, and the supplier of the test platform is working to ensure accessibility compliance in all parts of the system. The design of the student interface is based on the principle of universal design, which aims at creating solutions that work for as many users as possible from the very beginning of the design process.[21] The design of the system should be based on equivalence, flexibility, ease and intuitiveness of operation, perceivable information, tolerance for error, low physical effort, and size of access and use.

[18] Accessibility and usability are two different fields within user experience. Accessibility in the test platform focuses primarily on designing assignments in such a way that as many students as possible can do them, regardless of disability. Usability, on the other hand, refers to whether specified users can use the test platform to achieve their goals with a high degree of efficiency and satisfaction.

[19] The tests are designed at different universities. In all test development active teachers play an important role.

[20] WCAG 2.1 is a global standard specifically for web content accessibility. It provides guidelines and success criteria for making web content more accessible to people with disabilities.

[21] For more information, see this page on the Swedish Agency for Participation's website: https://www.mfd.se/verktyg/lar-om-politikens-mal-och-inriktning/politikens-fyra-arbetssatt/principen-om-universell-utformning/

The development of the student interface has been user-centered. During its development, Skolverket has applied different user-centered methods, and the interface has been tested by many students in different age groups, with different abilities, and in an authentic context. Insights from these tests have been fundamental in optimizing the usability of the test platform.

## Accessibility review of the exam content

In addition to an accessible and usable student interface, it is important that the actual content of the tests is also designed based on accessibility requirements and principles of universal design. The universities that Skolverket has commissioned to construct digital national tests have been responsible for designing test content in accordance with those requirements.

Skolverket supports the work of the universities to construct accessible tests in various ways. Skolverket developed a set of guidelines in the document "Design for accessibility 2.0", which the universities could use as reference material in their work. Since 2022, the content of each future digital national test has been reviewed by Skolverket based on accessibility requirements and principles of universal design. So far, Skolverket has carried out approximately 30 accessibility audits of digital national tests.

The purpose of reviewing the tests from an accessibility perspective is to ensure their accessibility and usability for all students to the greatest possible extent, so as not to limit the students' ability to demonstrate their knowledge in the test situation. On a more general basis, when Skolverket reviews a test, we check its quality, accuracy, and correctness to ensure it meets the established standards and criteria. In doing so Skolverket provides feedback and suggestions to the university to improve the test before its final version.

## Delays due to the Schrems ruling[22]

On 16 July 2020, the Court of Justice of the European Union delivered its judgment in the Schrems II case. Among other things, the ruling has had an impact on the test platform that Skolverket has procured for digital national tests since the system did not handle personal data in compliance with the General Data Protection Regulation (GDPR).

In 2021 and 2022, in response to the Schrems II ruling, Skolverket conducted a detailed and comprehensive review of the technical, legal, and operational aspects of the processing of personal data in the test platform. As a result, the provider of the test platform, which is based in Australia but owned by a company in the UK, has moved its operation and support of the test platform to the EU and the UK, respectively. The supplier has also undertaken to make changes to the test platform that are deemed to comply with the requirements of the EU's GDPR.[23]

[22] Schrems II, C-311/18, ECLI:EU:C:2020:559. The European Court of Justice's ruling Schrems II severely restricts the ability to transfer personal data to the United States from the EU and EEA.

[23] Government Review DNP 2023. P.24.

The consequences of the European Court of Justice's ruling in the Schrems II case[24] have had a considerable impact on the development of the test platform since the autumn of 2020. The supplier's relocation of operations and support from the USA and Australia to the EU and the UK, respectively, was completed in early 2023 (according to plan). The relocation of operations has delayed the development of the test platform as the supplier's capacity has been negatively affected.[25]

## Provisioning the users

Sweden lacks a centralized, national database that contains information about all school staff and students in the country. Therefore, all school organisers of schools in Sweden, as well as of Swedish schools outside the country, must provide Skolverket with information about their school staff and students. This process is called provisioning users.

In solutions for transferring data between the school organiser's technical infrastructure and the test service, Skolverket applies the information model and the value sets of the SS 12000 standard, "Interfaces for information exchange between operational processes in schools". The standard describes how each school organiser must collate their user data before transferring it to Skolverket's test service. This means that users in the physical world are represented digitally in the test service, as school staff or students, in a standardized way.[26]

In addition, unlike other countries, the Swedish school system lacks a centralized, national e-ID for all school staff and students.[27] It is the responsibility of each school organiser to ensure that an e-ID solution is implemented for school staff. From 2024 to 2026, Skolverket will provide a cost-free solution to school organisers who lack an e-ID solution. The solution is provided in cooperation with another agency, The Swedish Research Council, which has a government mission to adapt its product eduID[28] to the Swedish school system.

## Preparedness of schools and principals

As in most countries, the preparedness to use digital tools and communication rose during the COVID-19 pandemic. During the pandemic, it became more common for teachers to work in shared digital documents with their students, and for teachers to use digital test tools. There has also been a considerable increase in the use of digital learning materials in all stages of education. However, one major challenge in the implementation of digital e-testing at a national level is the variation in the digital preparedness of school organisers and their schools.

[24] Schrems II, C-311/18, ECLI:EU:C:2020:559. The European Court of Justice's ruling Schrems II severely restricts the ability to transfer personal data to the United States from the EU and EEA.

[25] Regeringsredovisning 2024. S. 18.

[26] DNP-redovisning 2024.

[27] Inloggningstjänst och e-legitimation - Skolverket.

[28] For more information, visit https://eduid.se/en/.

As mentioned above (in Background) the Swedish school system is decentralized and the responsibility for technical preparations lies largely on the school organisers and schools. This means that Skolverket has limited influence over the actions taken by school organisers and their schools when it comes to their preparations for digital national tests. Nevertheless. Skolverket has taken extensive steps to provide school organisers and schools with support material and information aimed at assisting them in their preparations for digital national tests.

### Support and communication efforts to school organisers, principals and schools

On Skolverket's website,[29] there is both information and support aimed at the target groups, especially school organisers, principals, teachers, and IT managers. Among other things, there is information about the preparations required for schools to be able to carry out digital national tests. This applies, for example, to the technical requirements that the school organisers and schools need to meet.

In addition to continuously providing information via the website, Skolverket uses social media and advertising to provide different groups with important information about digital national tests. There is also a targeted newsletter about digital national tests that is primarily aimed at school organisers, principals, and other levels of school management. In addition, Skolverket has published a newsletter that targets companies in the edtech sector that deliver digital services, products and IT systems to school organisers and schools. In addition, there is a guide on Skolverket's website for school staff who lead and organise work at the local level focused on digital national tests.[30]

Since Skolverket's test platform will be used for the first time in 2024, it is important that school staff become familiar with the system. Therefore, Skolverket offers an online training course that goes through the different parts of the test platform. Among other things, the course explains how school staff should prepare, carry out, and assess digital national tests in the test platform. The course was launched in January 2024.

In addition to web-based information and support, staff from Skolverket conducts webinars and participates in external events to inform about digital national tests and have direct discussions with different target groups.

Smaller school units face greater challenges to implement necessary requirements to run digital tests, e.g. federated login. Since grade 6 is the first group to use the test service in the spring of 2024, Skolverket has conducted webinars specifically aimed at school organisers who have a lower number of students in grade 6. Since the autumn of 2022, Skolverket has provided a support function that target groups can contact for technical and other questions related to digital national tests.[31]

[29] skolverket.se/dnp.

[30] For more information, see https://www.skolverket.se/om-oss/var-verksamhet/skolverkets-prioriterade-omraden/digitalisering/digitala-nationella-prov/guide-leda-arbetet-infor-digitala-nationella-prov.

[31] Government Report DNP 2024 pp 10-11.

### Challenges for school organisers and schools

The test service consists of several parts and the responsibilities are divided between Skolverket and the school organisers. A great deal of technical preparation must take place at both the school organiser level and at the school level. The technical preparations include:

- acquiring a technical solution that enables the provisioning of user information to Skolverket's test service
- implementing a technical solution for federated login in Skolverket's test platform
- obtaining e-ID for school staff to log in to Skolverket's test platform.
- ensuring that schools have digital devices and software that comply with requirements stipulated by Skolverket
- ensuring that schools have sufficient internet connectivity.[32]

During the past year Skolverket has visited schools, held discussions with school organiser, and conducted webinars and fairs. In connection to these activities, Skolverket has been made aware that several school organiser and schools have yet to begin technical preparations for digital national tests. It also appears that the knowledge of the organizational requirements for the implementation of digital national tests is still relatively low within schools.

This may be because school organiser and schools are awaiting Skolverket's formal regulations concerning digital national tests and decisions on several points that Skolverket and other authorities are investigating.

It may also be due to a lack of sense of urgency with school organisers and schools when it comes to the introduction of digital national tests. It appears as though routine activities have taken precedence, and digital national tests have not been given the appropriate level of priority. School organisers have also addressed a need for additional financial funding linked to digital national tests, but the government has not indicated that such specific funds will be made available.

To be able to carry out digital national tests in the best possible way, it is important that school organisers and schools make the necessary preparations and changes regarding technology and organization. These activities can be carried out as part of the systematic quality work that schools are to carry out regularly.[33] In our opinion, school organisers generally underestimate the time needed to prepare for the introduction of digital national tests. If school organisers and schools postpone necessary preparations, there is a considerable risk that students will not be able to take digital national tests in the test service. As mentioned above Skolverket will be providing replacement tests on paper for two years.

[32] Tekniska förberedelser och vägledning - Skolverket

[33] The Education Act contains requirements for systematic quality work. It also states that the quality work at school level must be carried out with the participation of teachers, preschool teachers, other staff and students. The principal is responsible for the work. The national curricula also include requirements for quality work. Source: Systematiskt kvalitetsarbete – så fungerar det

### Tests in the pilot project

In 2023, Skolverket carried out different activities with the pilot schools. Some examples are the processing of protected personal data in the population register, verification tests for login to the test service with e-ID, and the production of support material about digital national tests.

The most extensive activity was a technical test carried out in Skolverket's test platform in the autumn of 2023. All pilot schools were invited to perform so-called end-to-end tests of the test service. Skolverket offered an information meeting before the tests. Out of 99 pilot schools, 57 accepted and 42 declined to participate. The main reason for schools not participating was that the pilot school and its school organiser did not meet the technical requirements for digital national tests.

There can be several reasons why many of the pilot schools did not have the technical prerequisites for digital national tests in place. Some of them are listed below.

- Shortcomings in the cooperation between the school organiser and the principal.
- The principal awaits solutions for test administration.
- Lack of technical competence, either at the school organiser level or at the school level.
- Other, more urgent needs in the routine activities were given higher priority, such as the replacement of student registers.

Even though there was a small number of pilot schools that were able to participate in the end-to-end tests with successful results, Skolverket believes that the test as a whole was successful. The tests helped Skolverket identify several areas for improvement in the test service. The test results also show that school organizers and schools need more detailed information and support material connected to the introduction of digital national tests. This includes detailed information about provisioning user data, log in to the test platform, establishing clear division of responsibility between the school organiser and school, and other information that can assist the users of the system in preparations for digital national tests. Another insight from the tests is that Skolverket needs to optimize its support organization to ensure efficiency and accuracy. Also, it became clear that Skolverket has a new target group for digital national tests: IT unit or equivalent functions at the school organiser level or the school level.

## National assessment of digital national tests

The government mission concerning digital national tests was adjusted to include the implementation of a central assessment of the tests. Starting in the autumn of 2026, parts of the digital national tests will be assessed centrally instead of locally at the school in the following subjects: Swedish, Swedish as a second language, and English.

The aim is to make the assessment more fair and more equitable. When the system is fully deployed in 2028, 3,500 teachers will be working to assess approximately 400,000 student essays annually.

As mentioned above, external assessment has not been in use since the matriculation examination was abolished in the late 1960s and will therefore be a new phenomenon in Sweden.

To establish a central assessment, many actors, both within and outside the Swedish school system, will be faced with new challenges. Skolverket will play an integral role in launching the system internally and externally, attracting certified teachers who perform the assessment and thereby giving central assessment legitimacy.

A prerequisite for central assessment is that Skolverket's test service is up and running since the tests will be assessed digitally. The challenges involved will be to attract enough certified raters in time. To be eligible as a rater, the teacher will have to have the necessary certification in the subject being tested. The assessment will have to take place within a relatively short period of time, four weeks since the test result must be included by the teachers when setting the final grade of subjects and courses. The raters will assess an additional task, parallel to their regular work as teachers. The work as a rater will be compensated by Skolverket.

## Conclusions

To summarise this description of Skolverket's work with the implementation of national digital tests there are a few observations that stand out.

Firstly, it can be concluded that it has been a complex and multifaceted task to adapt the procured test platform to Swedish conditions. The body of requirements that is possible to describe in a procurement is dependent on the previous experience of in this case developing digital tests and the use of a national test service. The limited experience in defining requirements prolonged the process of adapting the platform for Skolverket's needs. In addition, the legal requirements in connection to the use of personal data changed during the implementation which also affected the pace in which Skolverket could proceed.

Secondly, in a decentralised education system like the one in Sweden the responsibility for digital national tests is divided between Skolverket and the school organisers. To ensure that schools and school organisers are adequately prepared Skolverket must understand the conditions of the target groups and give relevant information and support. Skolverket can stipulate some rules for the use of digital national tests but has limited possibilities to ensure that the whole process is in place at the right time. The agency can inform and support schools in their readiness and preparation.

The recent launch of the test service will show to what extent Skolverket has succeeded in this support.

Another aspect of the decentralisation is the mutual dependence between Skolverket, other agencies, schools, and school organisers. One example is the cooperation with The Swedish Research Council to establish a national e-ID-solution. Another is the need for school organisers to be prepared to supply the pupils' data.

## References

Eurydice (2024). A Decentralised Education System (online). Available: Overview (europa.eu) (March 2024).

Lundahl, C. (2009). *Varför nationella prov? – framväxt, dilemman, möjligheter.* Lund: Studentlitteratur.

Regeringen (2017). Uppdrag att digitalisera de nationella proven m.m. Available: Uppdrag att digitalisera de nationella proven m.m. - Regeringen.se. (March 2024).

Regeringen (2023). Likvärdiga betyg och meritvärden. Dir. 2023:95. Available: Likvärdiga betyg och meritvärden - Regeringen.se (June 2024).

Riksdagen. Förordning (2017:1106) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och central rättning. Available: Förordning (2017:1106) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och central rättning | Sveriges riksdag (riksdagen.se). (March 2024).

Skolinspektionen. Available: Inspektionsformer (skolinspektionen.se).(June 2024).

Skolverket (2014). *Privata aktörer inom förskola och skola - Skolverket*. Available: Privata aktörer inom förskola och skola - Skolverket. (June 2024).

Skolverket (2020). *Huvudmännens arbete med skolans digitalisering*. Rapport 2020:5. Available: Huvudmännens arbete med skolans digitalisering - Skolverket (March 2024).

Skolverket (2022). *Skolverkets uppföljning av digitaliseringsstrategin 2021*. Rapport 2022:4. Available: Skolverkets uppföljning av digitaliseringsstrategin 2021 - Skolverket. (March 2024).

Skolverket (2023). Redovisning DNP 2023. Available: pdf11127.pdf (skolverket.se) (June 2024).

Skolverket (2024). Redovisning DNP 2024. Available: Redovisning av uppdrag att digitalisera de nationella proven (skolverket.se). (June 2024).

Skolverket. Systematiskt kvalitetsarbete – så fungerar det Available: Systematiskt kvalitetsarbete – så fungerar det - Skolverket (March 2024).