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

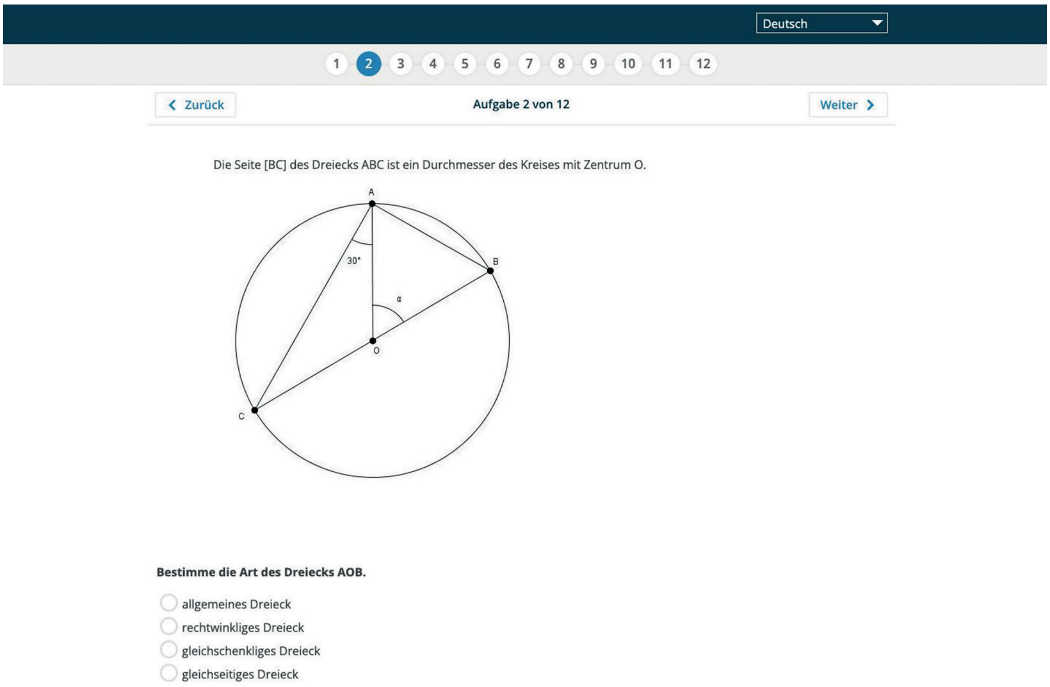
Luxembourg's response to e-assessment demands: the online assessment system (OASYS)

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.



Aufgabe 2 von 12

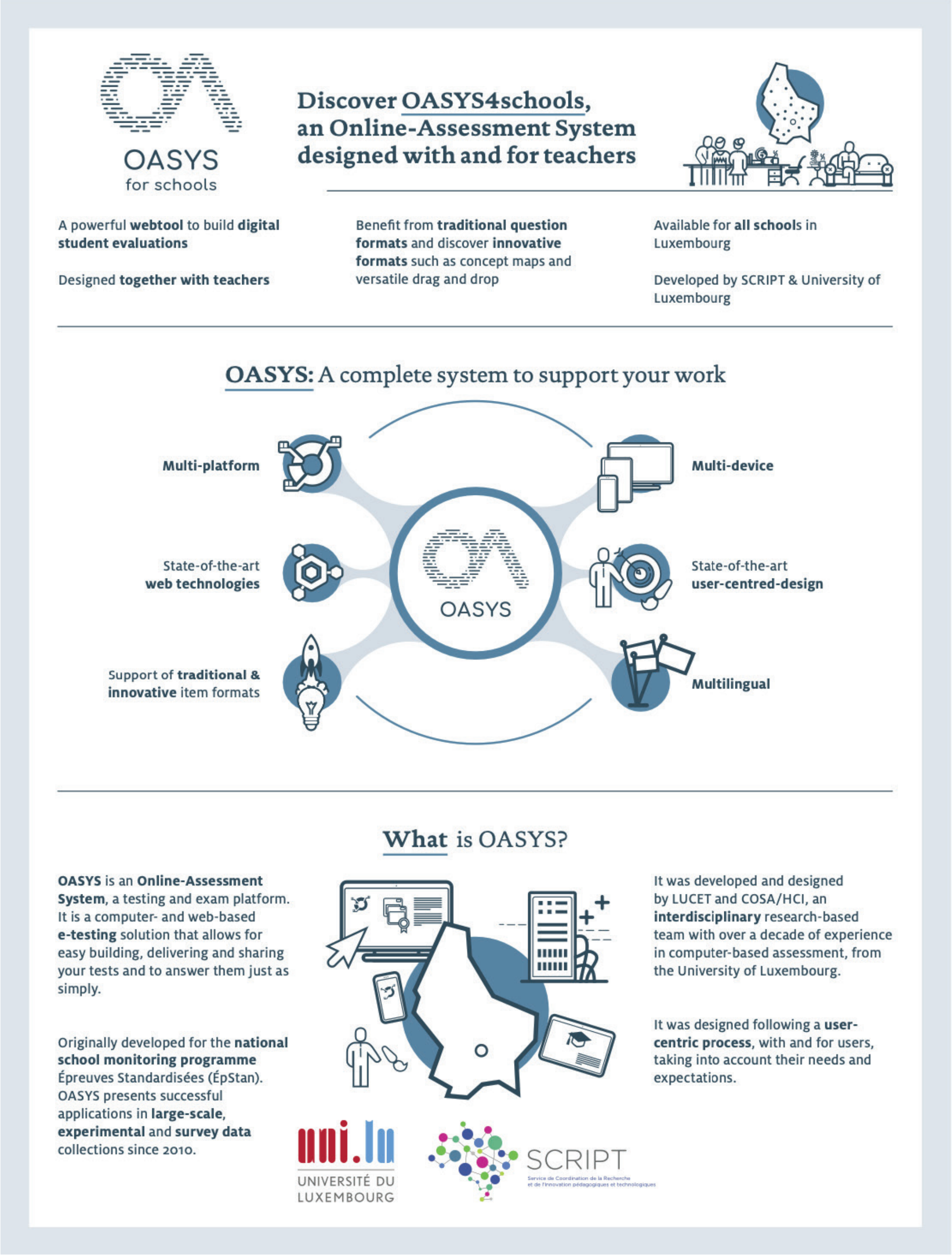
Die Seite [BC] des Dreiecks ABC ist ein Durchmesser des Kreises mit Zentrum O.

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

Multi-platform
Multi-device
OASYS
State-of-the-art web technologies
State-of-the-art user-centred-design
Support of traditional & innovative item formats
Multilingual

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.

uni.lu
UNIVERSITÉ DU LUXEMBOURG
SCRIPT
Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques

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).

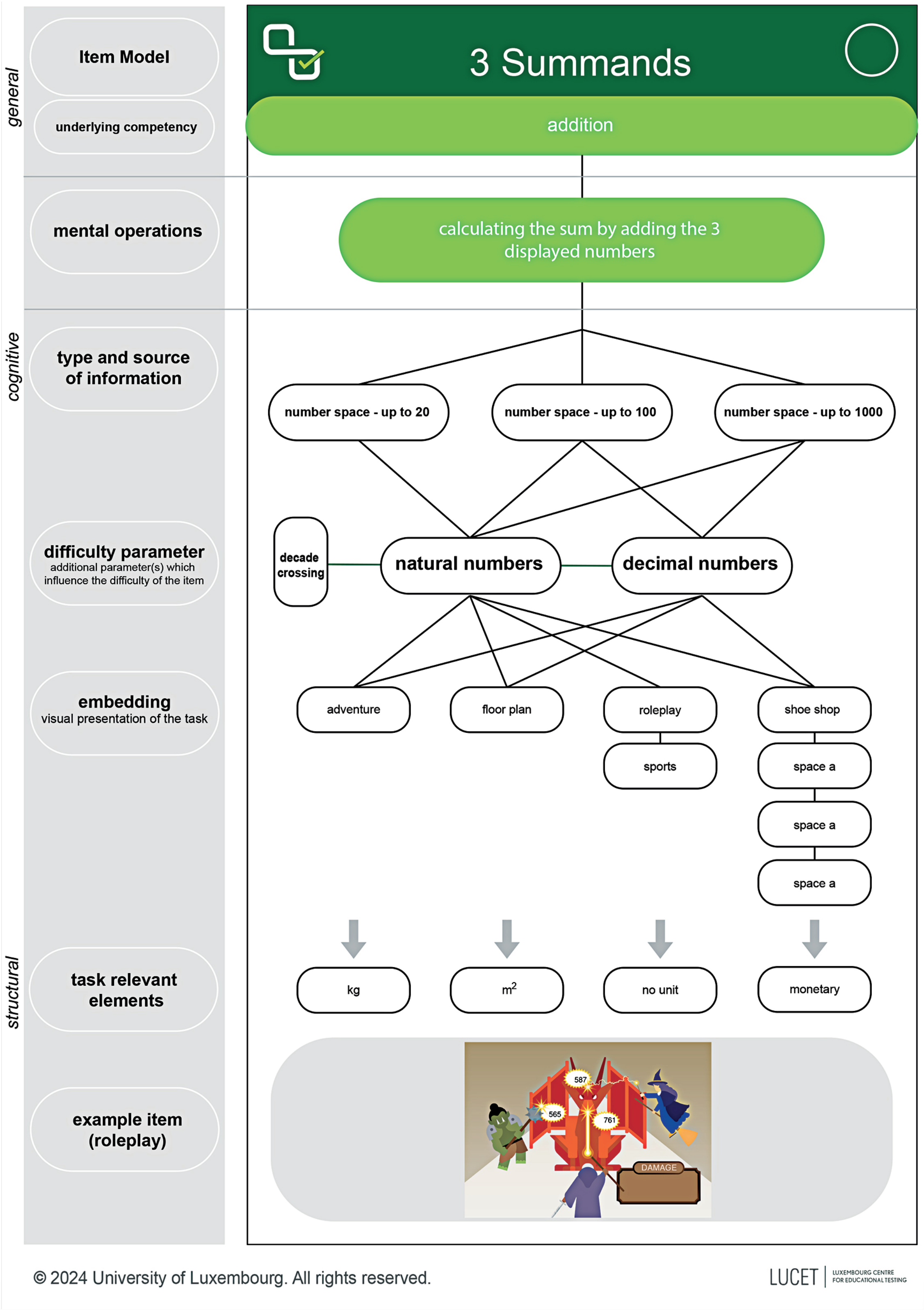


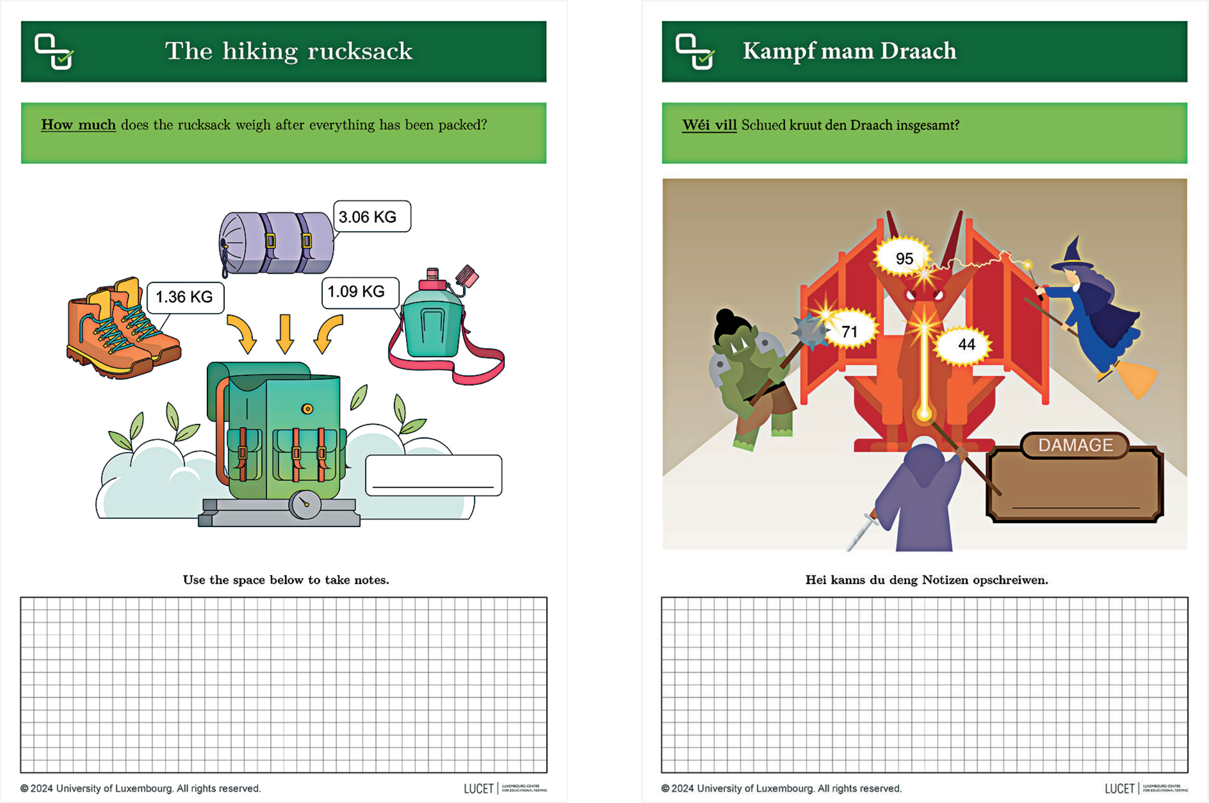
Figure 3: Cognitive model for the competency to calculate the sum by adding 3 numbers

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.

Kampf mam Draach

Wéi vill Schued kruut den Draach insgesamt?

Hei kanns du deng Notizen opschreiwen.

© 2024 University of Luxembourg. All rights reserved.

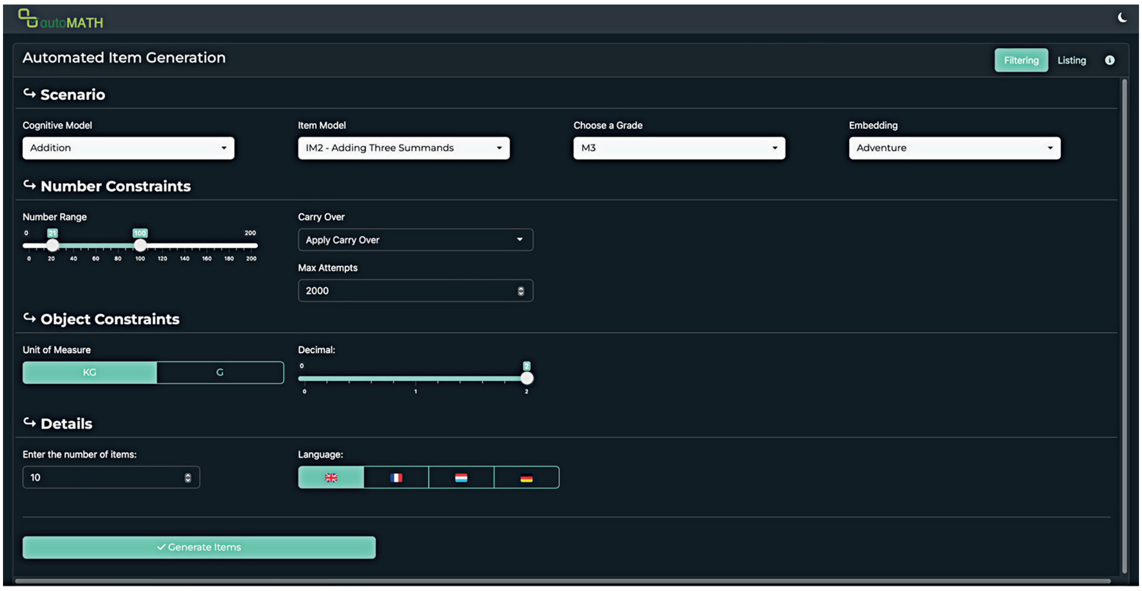
Figure 4: Two generated items for the competency of adding three summands including the cognitive constraints “decimal numbers” (left) and “decade crossing” (both). These examples also demonstrate the models’ language implementation in English, Luxembourgish, German, and French

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.



Automated Item Generation

Scenario

Cognitive Model: Addition

Item Model: IM2 - Adding Three Summands

Choose a Grade: M3

Embedding: Adventure

Number Constraints

Number Range: 21 – 100

Carry Over: Apply Carry Over

Max Attempts: 2000

Object Constraints

Unit of Measure: KG

Decimal: 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.

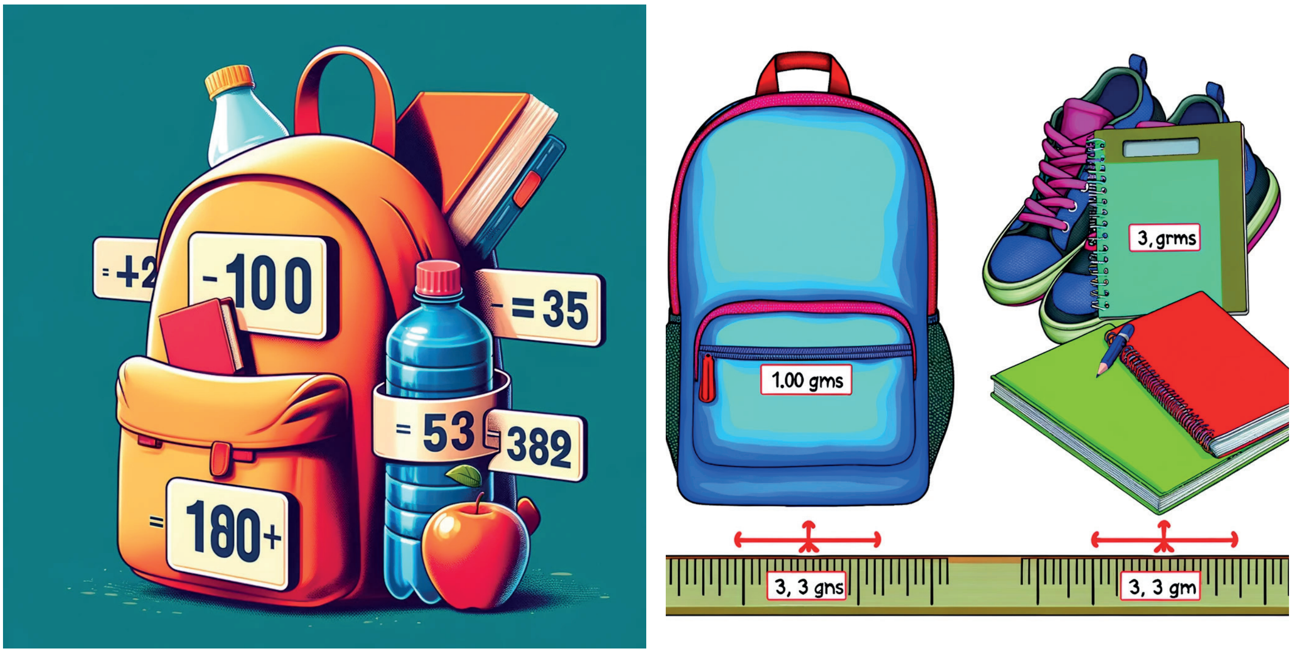


Figure 6: first two images generated by Dall-E/ChatGPT4 using the prompt "Draw me an image of a backpack. This backpack is being packed with three objects. Each object has a weight tag to it in grams. The total weight of the three objects should not exceed 1000g."

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.