IRSTI 14.91 https://doi.org/10.63597/UTO3105-4161.2025.4.4.001

N. Dieteren
Dutchtestologist, The Netherlands
e-mail: dutchtestologist@outlook.com

ORCID 0009-0009-7887-3745

THE ASSESSMENT OF HIGHER-ORDER THINKING SKILLS IN HIGH STAKES TESTING: AN UPDATED VISION ON HOW TO DEVELOP VALID AND RELIABLE ASSESSMENT INSTRUMENTS IN 21ST CENTURY CONTEXTS

Valid and reliable assessment of higher-order thinking skills (HOTS) in high-stakes tests requires a careful process of item writing. It starts with a sensible and well-reasoned choice of a taxonomy or framework, in combination with and based on a common understanding of the concept of higher-order thinking. Several taxonomies and a framework are presented as options to choose from. Each classification has its merits and should be seen as a guideline, not a law. The combination of the use of a taxonomy with the model of Cognitive Load Theory can be relevant for better understanding cognitive processes as operated by students when answering HOTS tasks in exams. Assessing HOTS in the 21st century requires the use of real-life contexts in exam tasks. Standards and criteria for the right choice and preparation of contexts are given and illustrated by some examples from high-stakes tests from the Netherlands and Kenya. Both open and closed item formats can be used for assessing HOTS in high-stakes tests. For all formats, it is important to devote attention and preparation time to the construction of valid keys and distractors (closed) or valid and operational marking schemes (open).

Keywords: HOTS, High-Stakes Testing, Validity, Reliability, Taxonomies, Assessment in Contexts.

Introduction

"In the last decades the meaning of knowing has shifted from being able to remember and reproduce information to being able to find, select, judge and use information in a productive way" [1]. In all communications and publications related to the PISA international survey on functional literacy, one can read that students should be prepared for life skills in 21st century contexts: more application, analysis, evaluation, creation [1]. These statements illustrate what kind of change is expected in 21st century education, as well as in the assessment of the learners' skills.

At the same time, assessments are still very important measurement instruments that are used worldwide to support essential decisions about learning progress and the development of individuals. For these high-stakes assessments, it is vital to realize very high standards in validity, reliability, and transparency of the assessment. Can we achieve these high standards and at the same time develop and implement more complex test items that reflect and demand higher-order thinking skills of the test candidates? And can we develop closed item types that combine objective or even automated scoring with the assessment of higher-order thinking skills in a high-stakes assessment instrument? These are the real challenges when it comes to assessment of higher-order thinking skills in 21st century contexts.

Materials and research methods

The use of taxonomies and definitions of higher-order thinking skills

In the 2016 paper, we focused on the assessment of the so-called 'higher-order thinking skills' (HOTS). The concept of 'higher-order thinking skills' (HOTS) is based on learning taxonomies such as Bloom's Taxonomy. In this taxonomy, skills that require processes of analyzing, evaluating, and creating are thought to be of a higher order, requiring different learning and teaching methods than the learning of facts and concepts. Higher-order thinking involves the learning of skills such as

critical thinking and problem solving. Higher-order thinking is thought to be more valuable because such skills are more likely to be usable in real-life situations (i.e., situations other than those in which the skill was learned).

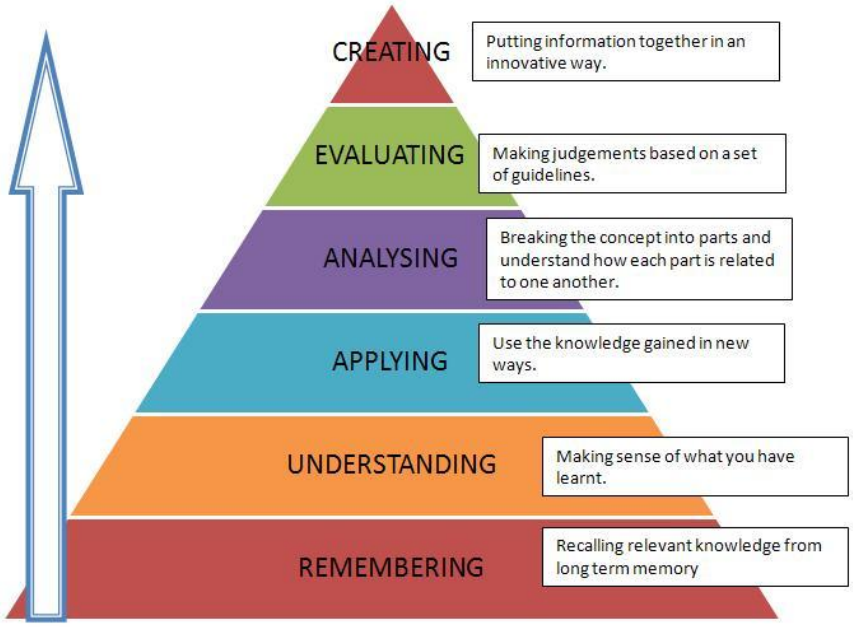


Figure 1 - Traditional representation of a hierarchical taxonomy
(Creative Commons licensed: https://creativecommons.org/licenses/by-sa/2.0)

More recently, the use of hierarchical taxonomies, like the revised Bloom's or SOLO [2], has been criticized by some scholars, as not fitting the real-life learning processes and mental operations by students and test takers [3, 4]. These scholars argue that Bloom's should not be viewed as levels or a hierarchy, but rather broken into lower-level and higher-level learning. Some find an inverse pyramid to be a better representation, with creating, evaluating, and analyzing at the top. One might argue that these three skills at the top can be defined as higher-order thinking skills.

In my practice as a test developer and teacher trainer for item writers, I mainly worked in countries where education was or is in the middle of a fundamental transformation process from knowledge-based teaching, learning, and assessing to competency-based education. For these contexts, higher-order thinking can already be observed when students are able to make productive use of the knowledge they learned in school, rather than merely reproduce what was taught. In such cases, I often classified all mental operations at applying level and higher as being higher-order thinking.

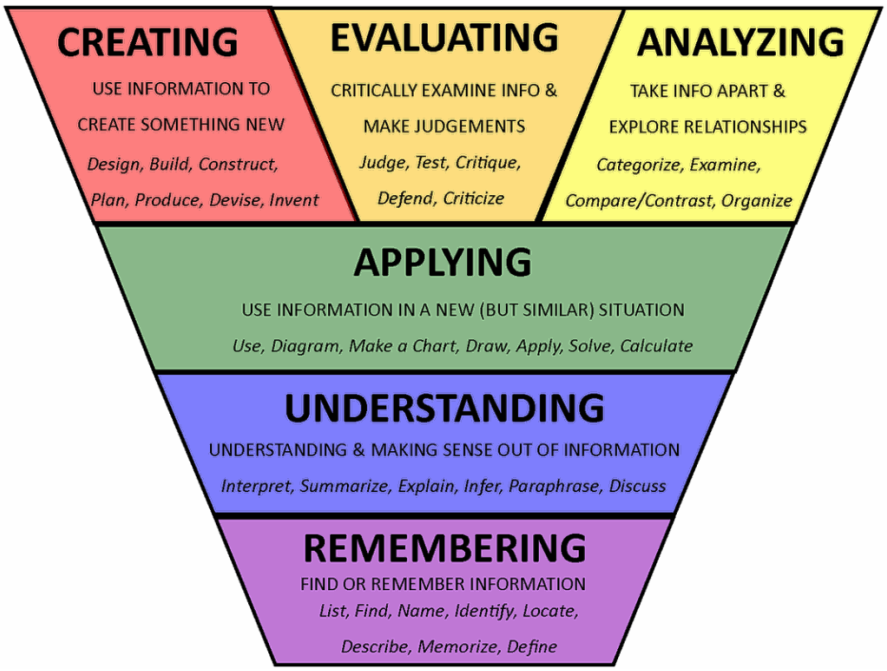


Figure 2 - Revised Bloom's taxonomy of learning processes, partly hierarchical
(www.intentionalcollegeteaching, 2021 [5])

Other scholars see additional nuances. In Anton Tolman's representation (see figure 3), the foundational skills of Understanding and Remembering are considered basic skills, while Analyzing, Creating, and Evaluating are considered critical thinking skills. Tolman sees Application as the transition or bridge that connects this necessary knowledge and more advanced 'higher-order' thinking skills [5, 6].

Personally, I very much like this approach, as it fits best with my training approach as described above. In transformational settings in education, where teachers and learners have to step out of their 'traditional habits' of rote learning and teaching to the test, the first hurdle to take is to get onto that bridge called 'Applying'.

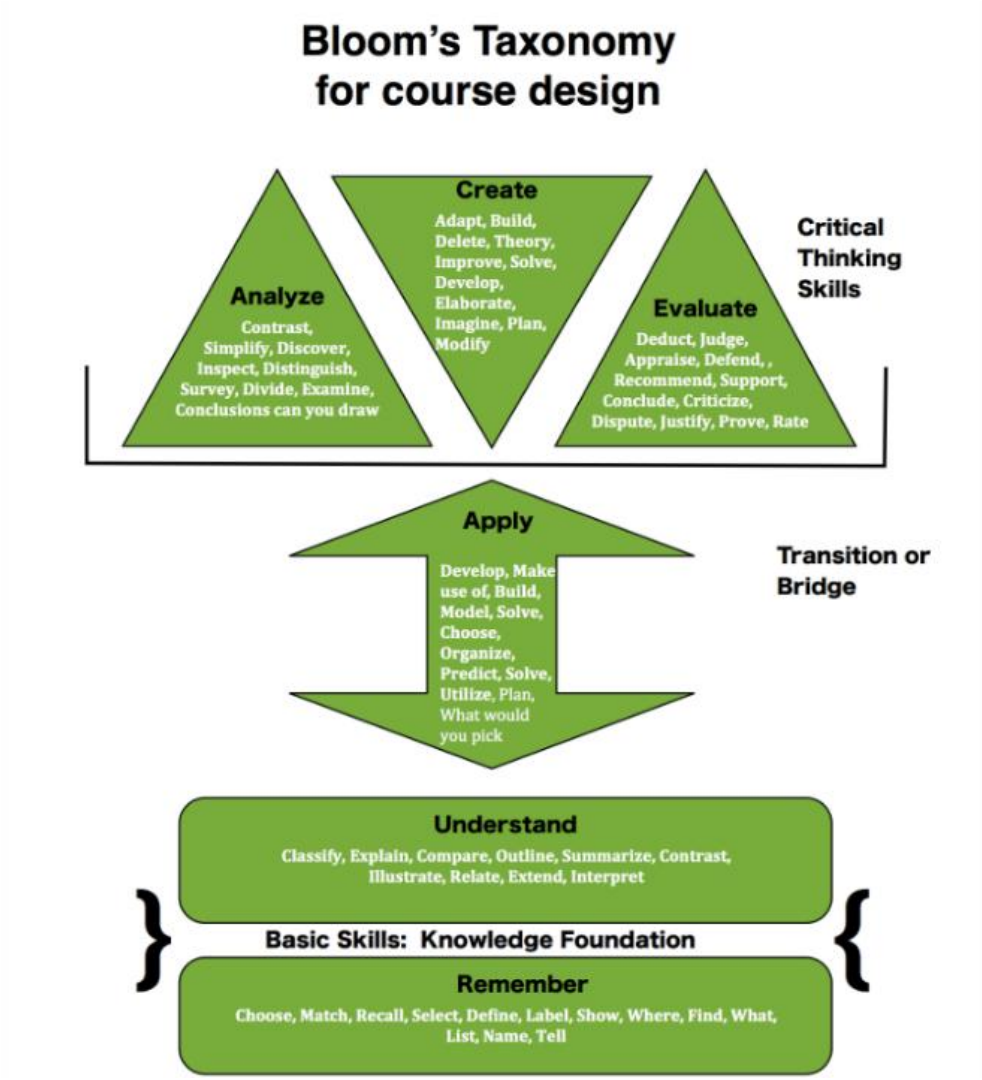


Figure 3 - Tolman's Cognitive Taxonomy Model
(www.intentionalcollegeteaching, 2021 [5])

An interesting and new perspective on the definition of higher-order thinking skills and the use of taxonomies for the development of HOTS tasks has been brought forward recently by Gulbakhyt Sultanova (CPI/NIS, Kazakhstan). She has performed a comprehensive theoretical study on the integration of two influential frameworks for defining and classifying tasks that assess higher-order skills in STEM education [1] [7]: the revised Bloom's taxonomy on cognitive processes and Sweller's theory on Cognitive Load. This second framework can be illustrated by the picture as shown below.

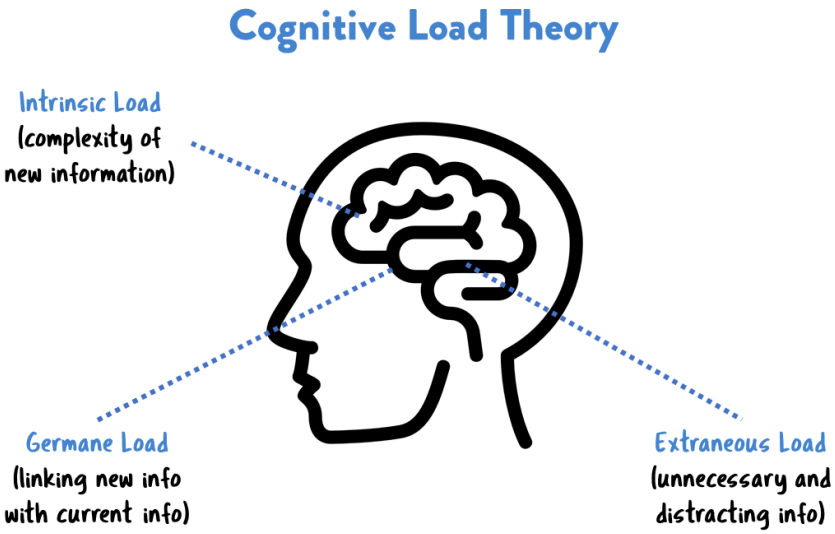


Figure 4 - A pictorial description of the Cognitive Load Theory
(www.barefootteflteacher.com; posted July 23, 2022)

Sultanova argues that the use of the insights of this Cognitive Load Theory (CLT) can improve the effectiveness of designing assessment tasks that aim to foster and display higher-order thinking skills [7]. In the research done by the 'founding father' of CLT, John Sweller [8], the cognitive processes related to problem-solving activities by learners were investigated by looking at how the working memory of learners functions in combination with long-term memory. In short, this looks like what is illustrated in Figure 5.

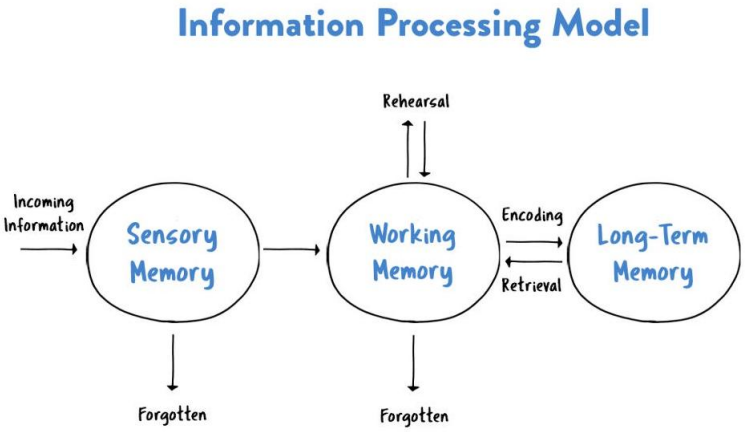


Figure 5 - A pictorial description of the Information Processing Model in CLT
(www.barefootteflteacher.com; posted July 23, 2022)

Imagine a student sitting an exam and being tasked to solve some items that demand higher-order thinking skills. By reading the text and stimulus material provided, the cognitive process starts with incoming information (left side in Figure 5) that has to be 'digested' by the working memory. What the developers of the exam task hope for is that the working memory will be able to rehearse some previously received/learned information that relates to this task, before the working memory sends all information for decoding by the long-term memory. In order to solve the task (hopefully successfully!), the working memory needs to retrieve all digested information and process this into an answer to the task. However, the working memory of the learner has a limited capacity.

Compare it to a computer where the RAM (Random Access Memory) has a limited capacity, whereas a hard drive is more like long-term memory.

According to the study by Sultanova [7], the germane load plays the central role in developing and using mental structures that help learners solve higher-order thinking tasks in scientific literacy. It is defensible to state that this goes for all higher-order thinking tasks that assess functional literacy in most common subjects in education [9]. More specifically, the germane load supports building schemas, being structured representations of knowledge stored in long-term memory. Building these structures repeatedly by doing such tasks in classroom practice and through regular coursework and assessments allows learners to build meaningful units. By 'filling' long-term memory with these units, the learner reduces the intrinsic load that is needed when being assessed with new higher-order tasks.

Summarized to the essential research hypotheses, as developed by Sultanova, the intrinsic load and the extrinsic load have a negative correlation with the results of learners when asked to demonstrate higher-order thinking skills in complex tasks. The amount and preparedness of germane load processed by the working memory, however, is supposed to have a positive correlation to these results on performing complex tasks. In the context of this paper, I will not go further into detail of this new integrative theory on the relation between higher-order thinking skills and the cognitive processes in the minds of the learners. The reader is referred to the full study when seeking more detailed information about the implications of this new theory for the development of effective, valid, and reliable tasks for assessing higher-order thinking skills in high-stakes tests [7].

An alternative approach for defining higher-order thinking skills

An alternative approach to the definition of higher-order thinking skills and the use of a system for classification of test items, be it a taxonomy or a framework, comes from the PACIER Framework, as developed by MACAT [10, 11].

	Problem Solving	**Analysis**	**Creative Thinking**	**Interpretation**	**Evaluation**	**Reasoning**
SKILLS	Producing strong solutions	Understanding how an argument is built	Creating new connections and unexpected solutions	Looking at issues of meaning	Exploring strengths and weaknesses of an argument	Creating strong and persuasive arguments
SUB-SKILLS	Asking productive questions	Working out the functions of each part of an argument	Connecting things together in a new way	Seeking to clarify meaning where necessary	Judging the acceptability of the reason(s) used in terms of readability	Producing well-structured arguments
	Generating possibilities	Understanding the relationships between parts of an argument	Producing novel explanations for existing evidence	Grasping the meaning of technical terms	Judging the relevance of the reason(s) used	Dealing with counter-arguments
	Generating solutions	Showing the structure of an argument	Generating new hypotheses	Understanding the meaning of available evidence	Judging the adequacy of the reason(s) used	Evaluating the reasoning of arguments
	Making sound decisions	Looking for assumptions in an argument	Redefining an issue so as to see it in a new way	Highlighting problems of definition	Judging what would strengthen or weaken an argument	Looking at the need to persuade

Figure 6 - The PACIER model for Critical Thinking

The PACIER model, developed by Macat in collaboration with the University of Cambridge, highlights six skills that collectively describe the propensity to think critically. When these related but somewhat distinct skills are combined, they become facets of critical thinking as a greater whole. Each letter of the acronym 'PACIER' represents one skill, and each skill breaks down into four sub-skills.

Compared to the revised Bloom's taxonomy, the concept of higher-order thinking skills is broadened by the PACIER model with the addition of problem-solving, interpretation, and reasoning. To cover these six main skills, the concept of Critical Thinking is used as an umbrella concept.

Figure 6 specifies each of the six skills and lists associated sub-skills. With respect to the American Philosophical Association's consensual definition, self-regulation relates to problem solving; analysis, interpretation, and evaluation to the same components in PACIER; and inference to reasoning.

For clarity of reading and unity in understanding, I will use the term higher-order thinking skills (sometimes abbreviated as HOTS) throughout the rest of this article. The reader should be aware of the diversity of sub-skills that can be part of this 'container concept', as is most comprehensively described by the PACIER framework.

HOTS or Critical thinking skills and the relation to assessment in contexts

There are many aspects to be addressed regarding assessment of higher-order thinking skills. An important aspect is the use of contexts in the assessment. The way in which real-life contextual situations are represented depends on the delivery form of the assessment. This may be paper, computer, simulator, or practice. How can we assess higher-order thinking skills by assessment in contexts?

In this contribution, we show advantages and pitfalls of assessment in contexts: assessing knowledge and skills in a given situation. How can we best select suitable contexts for the various types of assessment? How can we construct tasks with these contexts? And how to construct reliable and objective marking schemes when these tasks are open-ended questions?

In the previous chapters, we have seen that there are many frameworks that give a definition and describe how higher-order thinking skills are learned and developed. Higher-order thinking involves a variety of cognitive processes applied to complex situations. The specific cognitive process to be applied depends on the actual real-life situation at stake: the so-called 'context'. Assessment in contexts is assessing knowledge and skills in real-world situations, but assessment usually does not take place in the real world. Depending on the delivery form of the assessment – paper, computer, simulator, practice – the presentation of the context is an approximation of the real world. In paper-based tests, the context will be described and shown or illustrated by pictures, charts, and tables: the so-called stimulus material. In computer-based tests, the context can be designed in more variety and sometimes more realistically, especially when videos or simulations can be used. Take, for example, a flight simulator test that is often used in the assessment of trained pilots. This test is a very close approximation to real life. In some situations, the education or training can be finalized with a real practical test. In that case, the context is actually the real-life situation. A very well-known example is the driving licence practical exam. But for most cases and subjects in general education, the assessments are predominantly paper-based or computer-based tests of theoretical cognitive knowledge. In this article, I will focus in more detail on these formats, which present a constructed model of the real-life situation. I will pay attention to several aspects of the contextual situation in the development of assessments. How can we select a suitable resource? Which quality standards should be applied? Have guidelines been developed that can be of help in constructing assessment tasks? How to assure objective and justifiable marking and scoring of assessment tasks made by candidates?

Standards and criteria for assessment in contexts

Assessment in contexts will always be related to a certain learning content or program. What is it that we want to observe when assessing the learners? The content of this learning can be defined in several ways. In general we see three types of learning content:

- Curricula, as mostly developed and implemented for all relevant school subjects; for example learning outcomes for mathematics in grade 9;
- Skills that are defined for certain domains of functional literacy; for example the PISA frameworks for reading, mathematical or scientific literacy [12];
- Competencies in job-related contexts; for example the skills needed to become a certified aviator.

Assessments in contexts will differ in structure, ways of displaying the context, types of tasks, and methods for marking and scoring, related to the three types of standard as indicated above. But for all, we state that a certain general list of criteria can be applicable when defining the quality and effectiveness of the assessment context to be chosen and developed. The list below shows an example for mathematics, natural sciences, and social sciences as learning subjects and also for mathematical and scientific literacy in skills-based surveys:

1 Situation should be realistic, but not necessarily real
2 Choose authentic materials if available
3 Stimulus materials must be free of copyrights and intellectual property rights
4 Context must be functional
5 Context must be natural
6 Context must be efficient
7 Context must be subject related
8 Context must be neutral
9 Context must be acceptable for all candidates
10 Contexts must correspond with the circle of interests of the candidates

This list of criteria is presented for the purpose of discussion with experts in the field, not necessarily to present the only way to select context materials. A comparable, but separate, list of quality criteria could be made for text choice in reading assessment.

With regard to the validity of the assessment in contexts, I have to emphasize that

- assessing in contexts is only valid when teaching, training and learning are done in contexts: text books and other learning materials, exercises and intermediate ‘observations’ should already train the learner in the productive and creative use of knowledge in different real-life situations;
- the test developer should always keep in mind that it is not the context that is assessed, but the knowledge and skills that are purposely related to that context.

Guidelines to the development of HOTS tasks

The development of tasks to assess higher-order thinking skills, is generally done in two steps. The first step is the development of the context, the second step is the development of items (see figure 7).

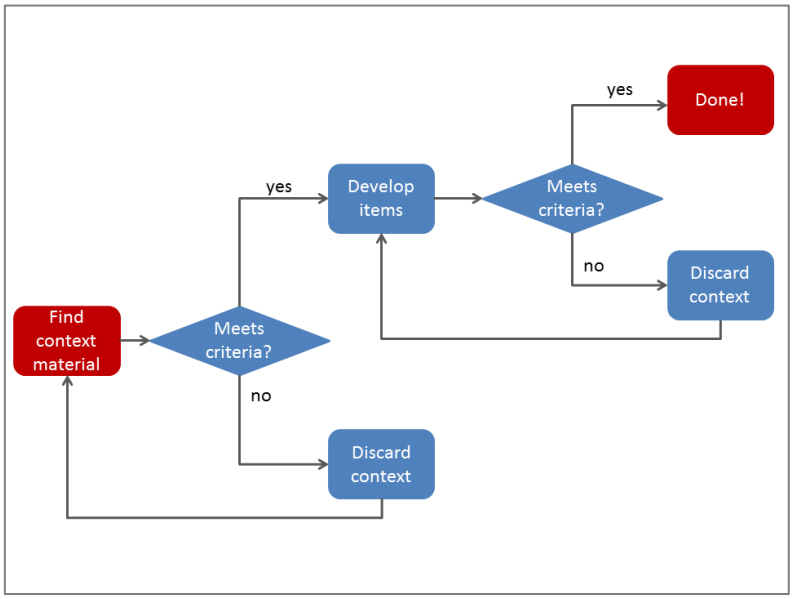


Figure 7 - Process of task development, assessment in context
© dutchtestologist

Based on our best practice, we made the following list of criteria for the development of items in context. An item must be:

1 relevant
2 at intended level
3 specific
4 objective
5 acceptable
6 transparent
7 efficient
8 in correct language
9 in clear layout

The translation of these criteria to their respective contribution to the validity and reliability of a test is not elaborated in full here. Such elaboration, including exercising and discussing with peers, is part of training courses that I used to provide for item writers and test developers.

In paragraph 7, three examples will be given of items/tasks that assess HOTS in context. Two from the Dutch final examinations in upper secondary education and one from a pilot project on Competency-Based Assessment (CBA) I did in Kenya with the teams from Cito (Netherlands) and KNEC (Kenya).

Results and Discussion

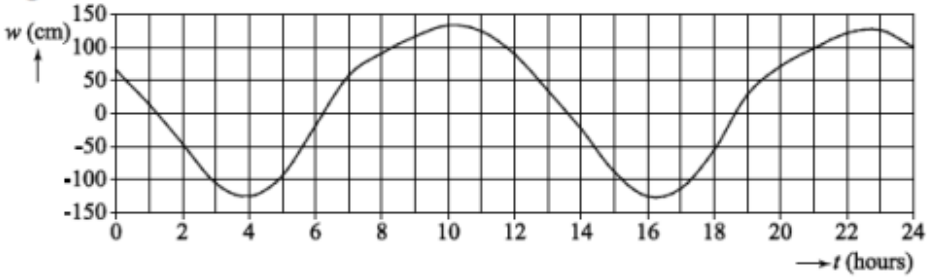
Examples of test items for assessing HOTS in contexts

The first example comes from an exam in Mathematics on the level of upper-secondary school leaving exams that are also used for general entrance qualification to tertiary education (universities). The level of the subject Mathematics can be compared with UNT level in Kazakhstan. The task comes from a paper-based exam and is translated from Dutch language to English for the purpose of sharing on international platforms.

High and low tide

Rijkswaterstaat publishes tide predictions for a number of places along the Dutch coast. These are calculated using a mathematical model based on measurements over a long period of time. Figure 1 gives the expected water level on 14 November 2012 for Schiermonnikoog.

figure 1



The values in the graph of figure 1 can be approximated by the formula $w = 4 + 128\sin(0.51(t + 5.4))$. Here, w is the water level in cm and t is the time in hours with $t = 0$ at 00:00. The moment of highest water level in the evening according to the formula differs from that moment in the graph in figure 1.

4p 17 Determine this difference in minutes.

Figure 8 - A HOTS task example mathematics from the Netherlands

Marking scheme
maximum score 4

- *Calculating the correct maximum of the graph results in $t \approx 22.3$* 1
- *Correct interpretation of the graph: in the evening high tide occurs at 22:40* 1
- *$t \approx 22.3$ corresponds to 22:18 (or 22:19)* 1
- *The difference is 22 (or 21) (minutes)* 1

Note:
A margin of 10 minutes is allowed in the time of high tide that has been written down.

Source: VWO Mathematics A 2014 1st term, Mathematics secondary school-leaving exam for upper secondary education in the Netherlands. © dutchtestologist

Note that the task given is highly complex and stated very openly, like a research question. The structure and composition of the marking scheme shows that the cognitive process that is expected from the student contains several operations and choices to be made. Each of these steps can be awarded 1 mark, up to a total of 4 for a fully correct answer.

Item writers who constructed such HOTS items for the final exams in Mathematics in the Netherlands used to work with an adapted Bloom's taxonomy, made collectively by the members of the Mathematics construction team. In the case of this item, they classified the task as being on the higher-order level of 'Analysing'. In the 2014 exam, this item #17 was the first in a cluster of 4 items, all related to the context material as shown above.

A comparable adapted Bloom's taxonomy table, as used recently in one of my international training and consultancy projects, is shown in the table below.

Table 1 - An adapted taxonomy for practical use in item writing Mathematics

Category of conceptual knowledge	Description(s)
Remembering (no context)	- Recall formulas, rules, terms - Recall calculations that should be automated
Understanding (no context)	- Identify and use calculations or operations, not put in a context - No need to choose/decide which operation to use
Applying (in a context)	- Use math tools, concepts and operations in a given (new) context - Procedure to use is not directly stated in the question/task - Mostly only one procedures needed (primary), sometimes more procedures needed (secondary)
Analysing (in a context)	- Use more than one operation/procedure/calculation along with some decision about which order to use them - Demands thinking about a strategy for answering
Evaluating (in a context)	- Several decisions to be made about which operations/procedures/calculations to use - Demands providing some conclusion or decision
© dutchtestologist	

For the sake of clarity and simplicity in practical use, the categories are listed top-down, not suggesting a 1 to 5 hierarchy. In this item-writing team, the experts decided to make a clear distinction between test items WITHOUT context and WITH context. Whenever using a context, like the contexts I described in paragraphs 5 and 6, the item(s) to be developed must aim at assessing a higher-order skill like applying, analysing, or evaluating.

The category of 'creating' was judged as 'not relevant' in this project, as all test items had to be closed formats and usable both for paper-based and computer-based assessments.

A second example comes from a computer-based exam for science in a vocational track (TVET) in the Netherlands. The task consists of a simulation, where the student has to close an electrical circuit using different elements. The elements can be dragged and dropped to the board on the right side, in accordance with the actual task and description of the situation. There are 3 open places where a student can place an element. In the end, the completed circuit must enable the measurement of the voltage of an LED light that is connected.

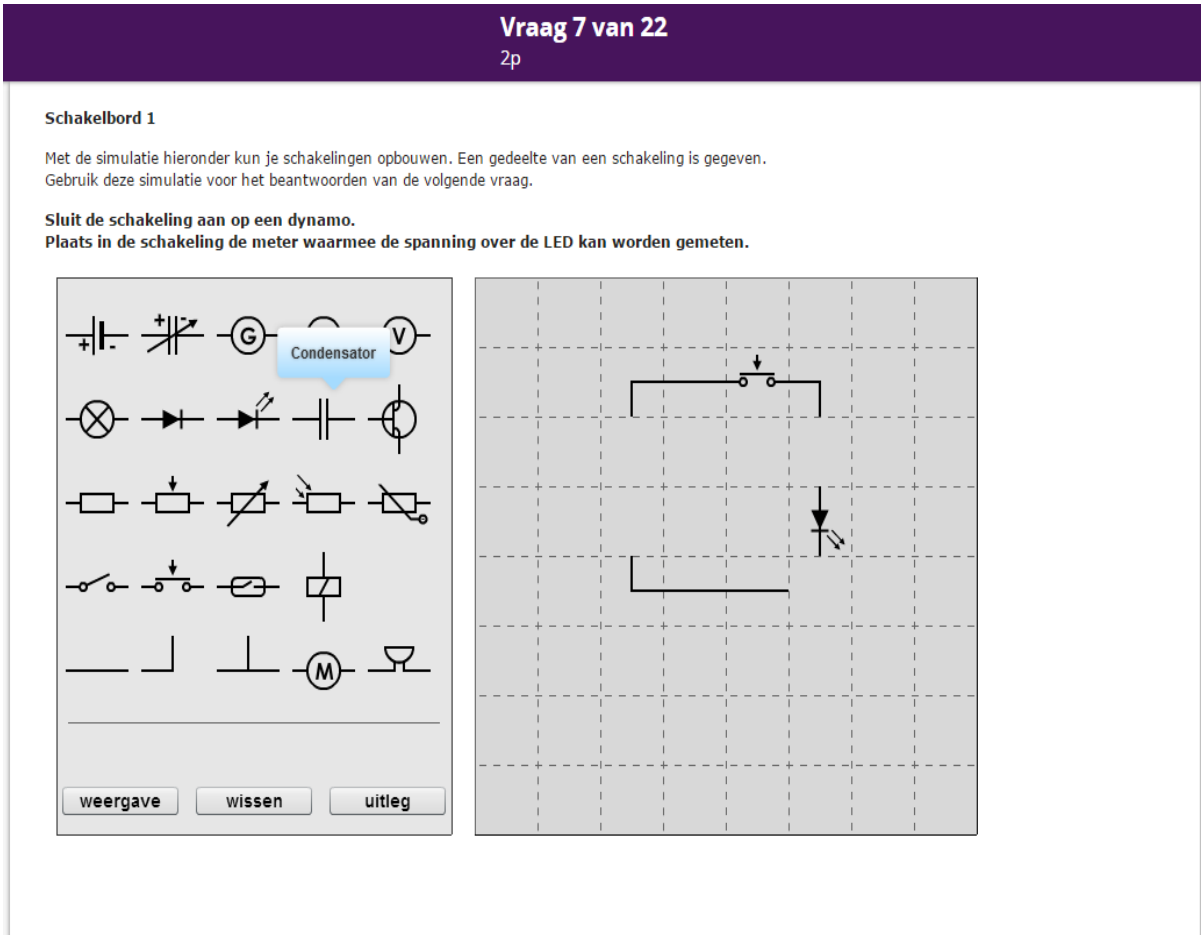


Figure 9 - A HOTS task example science from the Netherlands

Source: VMBO Physics 1 example, Physics secondary school-leaving exam in the Netherlands. Retrieved 1 July 2016 from https://vo-oefenomgeving.facet.onl

This exam task was classified on the level of 'applying' (secondary education).

One of the additional advantages of such computer-based items is that the marking can be done automatically by the software and a built-in analysis tool. Even partial credit can be assigned to certain choices that are not correct but could be reasonable in comparable settings. The item as such is a fine example of how the assessment in closed format of higher-order thinking skills in a high-stakes exam can be combined with very reliable and objective marking and scoring. I will elaborate on this topic in the next paragraph.

A third example comes from a pilot test for primary education students in Kenya. This task shows that assessment of higher-order thinking skills does not always demand complex tasks with elaborated stimulus material. In this case, a very basic level of applying is assessed, using what the student learned in science class about the concept of force.

13. The picture shows a learner demonstrating a certain effect of force on an object.

The demonstration shows that force

A. changes direction of an object
B. makes an object to start moving
C. changes shape of an object
D. makes an object to stop moving

Figure 10 - A HOTS task example from Kenya
Source: Kenya Primary School Education Assessment, Integrated Science, KNEC 2021

Importance of objective and unambiguous marking schemes for assessment in contexts

When indicating 'Done!' in the chart of Figure 7 above (paragraph 6), the work of the test developer is not finished. The development and implementation of a good marking scheme is an essential next step to assure a reliable and objective measurement of higher-order thinking skills. In fact, the correct key to a closed question or the correct marking scheme for open-ended questions is mostly seen as an integral and necessary part of a complete test item. Nevertheless, I want to pay special attention to that part, as in many cases test and item developers tend to focus much attention and development time on good questions or tasks and less on the quality and correctness of the answer models. The same can be said about some generative AI models, like ChatGPT or DeepSeek, when used for item development [13–15].

For assessing higher-order thinking skills, test developers mostly make use of open-ended item types. Candidates get more opportunities to show their productive and creative skills when allowed to give longer answers on open questions or tasks. But in high-stakes assessments, we place great value on objective marking and scoring of all answers given by different learners on the same tasks or questions. We want to be sure that equivalent ability of learners is marked and scored as equally as relevant. Therefore, marking schemes have to be unambiguous and clear. In most high-stakes assessments on a national scale, the works of candidates are marked and scored by a great number of raters. The personal view of the rater should not influence the candidates' results in the end. High standards for rater agreement are demanded.

In the case of more complex HOTS tasks, this aim of high rater agreement sets very ambitious standards for the marking schemes [16]. In general, we can state that most criteria for good items, as listed in the previous paragraph, also must be related to the marking schemes. In this paper, we illustrate, with two examples, the way marking schemes are developed according to these requirements.

In the case of closed item types, a marking scheme is already defined by the structure and the content of the task. Much of the development challenge for multiple-choice items lies in the correct formulation of the alternatives, including functional distractors. Or in the case of constructed-response items, the big issue is to make correct and unambiguous response restrictions. It is often stated that closed item types, especially the standard multiple-choice questions with 3 or more alternatives, are not very appropriate for assessing higher-order thinking skills [17].But, several

studies, mainly done in the world of high-stakes assessments in medical studies in higher education, have shown that there is evidence for reliable and valid assessments of higher-order thinking skills with the use of MCQ items [17-19]. These findings are in line with my own experiences, most successfully when assessing Bloom's skills like applying, analysing, and evaluating or PACIER framework skills like interpretation and problem solving.

Conclusion

This article presents some issues chosen from proven best practice in the development of assessments in context. If exams that aim at assessing higher-order thinking skills in a high-stakes situation are developed along the guidelines presented, it is believed that they contribute to better preparation of secondary school students for higher education specifically and for professional development in life in general.

Prenotice: this article is a revised and expanded version of a full paper, as written and presented for the 42nd Annual Conference of the International Association for Educational Assessment (IAEA), in Cape Town, South-Africa [20]

References

1. OECD. Programme for International Student Assessment (PISA). – 2024. - https://www.oecd.org/en/about/programmes/pisa.html
2. Biggs J. B., Collis K. F. Evaluating the Quality of Learning: the SOLO taxonomy. – New York : Academic Press. – 2014.
3. Case R. The Unfortunate Consequences of Bloom's Taxonomy // Social Education. – 2013. – Vol. 77 (4). – P. 196–200.
4. Soozandehfar S., Adeli M. A Critical Appraisal of Bloom's Taxonomy // American Research Journal of English and Literature. – 2016. – Vol. 2. – P. 1–9. – DOI:10.21694/2378-9026.16014
5. Tolman A. O., Kremling J. Why students resist learning: A practical model for understanding and helping students. – Stylus Publishing, 2017.
6. Bloom's Taxonomy: Benefits and Limitations // Intentional College Teaching. – 2021. – 30 April. - https://intentionalcollegeteaching.org/2021/04/30/blooms-taxonomy-benefits-and-limitations
7. Sultanova G. Assessment tasks in STEM education: integrating Bloom's Revised Taxonomy and Cognitive Load Theory // Research in Science & Technology Education. – 2025. - 10 Nov. – P. 1–18. – https://doi.org/10.1080/02635143.2025.2586564
8. Sweller J. Cognitive Load During Problem Solving: Effects on Learning // Cognitive Science. – 1988. – Vol. 12, № 2. – P. 257–285. – https://doi.org/10.1207/s15516709cog1202_4
9. Arwood E. L. An Introduction to Pragmatic Assessment and Intervention for Higher Order Thinking and Better Literacy. – UK : Jessica Kingsley Publishers, 2011. – 304 p.
10. Macat. – 2026. - https://www.macat.com/
11. Shin H. J., Li S., Ryoo J. H., von Davier A., Lubart T., Khalil S. The Nature and Measure of Critical Thinking: The PACIER Framework and Assessment // Journal of Intelligence. – 2025. – Vol. 13. – P. 1-16. – https://doi.org/10.3390/jintelligence13090113
12. OECD. PISA 2015 Draft Mathematics Framework. – March 2013. - https://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm
13. Leigh A. AI Question Generation: The Risks and Alternatives // Grademaker. – 2024. – 5 March. - https://www.grademaker.com/news/ai-question-generation-the-risks-and-alternatives/

14. Corbin T., Bearman M., Boud D., Dawson P. Assessment & Evaluation in Higher education: the wicked problem of AI and assessment // Assessment & Evaluation in Higher Education. – 2025. – P. 1–13. – https://doi.org/10.1080/02602938.2025.2553340

15. Kahn S. M., Hamer J., Almeida T. Generate: A NLG system for educational content creation // Proceedings of the 14th International Conference on Educational Data Mining. – 2021.

16. Black B., Bramley T., Suto I. The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement // Assessment in Education: Principles, Policy and Practice. – 2011. – Vol. 18 (3). – P. 295–318. – https://doi.org/10.1080/0969594X.2011.555328

17. Liu Q., Wald N., Daskon C., Harland T. Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers // Innovations in Education and Teaching International. – 2024. – Vol. 61 (4). – P. 802–814. – https://doi.org/10.1080/14703297.2023.2222715

18. Bibler Zaidi N. L., Grob K. L., Monrad S. M., Kurtz J. B., Tai A., Ahmed A. Z., Gruppen L. D., Santen S. A. Pushing Critical Thinking Skills With Multiple-Choice Questions: Does Bloom's Taxonomy Work? // Academic Medicine. – 2018. – Vol. 93 (6). – P. 856–859. – https://doi.org/10.1097/ACM.0000000000002087

19. Stringer J. K., Santen S. A., Lee E., Rawls M., Bailey J., Richards A., Perera R. A., Biskobing D. Examining Bloom's Taxonomy in Multiple Choice Questions: Students' Approach to Questions // Medical Science Educator. – 2021. – Vol. 31, № 4. – P. 1311–1317. – https://doi.org/10.1007/s40670-021-01305-y.

20. Dieteren N., Jongkamp C. Higher-order thinking skills in high stakes testing: how to develop valid and reliable assessment instruments in 21st century contexts? // 42nd Annual Conference of the IAEA. – Cape Town. – 2016.

Н. Дитерен

ЖОҒАРЫ ДӘРЕЖЕЛІ ОЙЛАУ ДАҒДЫЛАРЫН ЖОҒАРЫ МАҢЫЗДЫ ТЕСТІЛЕУДЕ БАҒАЛАУ: 21-ШІ ҒАСЫР КОНТЕКСТТЕРІНДЕ ВАЛИДТІ ЖӘНЕ СЕНІМДІ БАҒАЛАУ ҚҰРАЛДАРЫН ӘЗІРЛЕУДІҢ ЖАҢАРТЫЛҒАН КӨРІНІСІ

Жоғары маңызды тесттерде жоғары дәрежелі ойлау дағдыларын (ЖДОД) валидті және сенімді бағалау тапсырмаларды мұқият әзірлеуді талап етеді. Бұл жоғары дәрежелі ойлау тұжырымдамасын ортақ түсінуге негізделген және онымен біріккен таксономияны немесе құрылымды орынды және дәлелді таңдаудан басталады. Таңдау нұсқалары ретінде бірнеше таксономия және бір құрылым ұсынылған. Әрбір жіктеудің өзіндік артықшылықтары бар және ол заң емес, нұсқаулық ретінде қарастырылуы тиіс. Таксономияны Когнитивтік жүктеме теориясы (КЖТ) моделімен біріктіру емтихандарда ЖДОД тапсырмаларына жауап беру кезінде оқушылардың когнитивтік процестерін жақсы түсіну үшін маңызды болуы мүмкін. 21-ші ғасырда ЖДОД-ны бағалау емтихан тапсырмаларында нақты өмірлік контексттерді пайдалануды талап етеді. Контексттерді дұрыс таңдау және дайындау үшін стандарттар мен критерийлер берілген және олар Нидерланды мен Кенияның жоғары маңызды тесттерінен алынған мысалдармен суреттелген. Жоғары маңызды тесттерде ЖДОД-ны бағалау үшін ашық және жабық тапсырма форматтарын да пайдалануға болады. Барлық форматтар үшін валидті кілттер мен дистракторларды (жабық) немесе валидті және жұмыс істейтін бағалау схемаларын (ашық) құруға назар аудару және дайындық уақытын бөлу маңызды.

Түйін сөздер: ЖДОД, жоғары маңызды тестілеу, валидтілік, сенімділік, таксономиялар, контексттерде бағалау.

Н. Дитерен

ОЦЕНИВАНИЕ НАВЫКОВ МЫШЛЕНИЯ ВЫСШЕГО ПОРЯДКА В ВЫСОКОЗНАЧИМОМ ТЕСТИРОВАНИИ: ОБНОВЛЕННОЕ ВИДЕНИЕ ТОГО, КАК РАЗРАБАТЫВАТЬ ВАЛИДНЫЕ И НАДЕЖНЫЕ ИНСТРУМЕНТЫ ОЦЕНИВАНИЯ В КОНТЕКСТАХ 21-ГО ВЕКА

Валидная и надежная оценка навыков мышления высокого порядка (НМВП) в тестах с высокими ставками требует тщательного процесса составления заданий. Она начинается с обоснованного и взвешенного выбора таксономии или концептуальной рамки в сочетании с общим пониманием концепции мышления высокого порядка и на его основе. Представлены несколько таксономий и концептуальная рамка в качестве возможных вариантов для выбора. Каждая классификация имеет свои достоинства и должна рассматриваться как ориентир, а не как закон. Сочетание использования таксономии с моделью теории когнитивной нагрузки может быть полезным для лучшего понимания когнитивных процессов, задействованных учащимися при выполнении заданий на НМВП на экзаменах. Оценка НМВП в XXI веке требует использования реальных жизненных контекстов в экзаменационных заданиях. Стандарты и критерии правильного выбора и подготовки контекстов представлены и проиллюстрированы примерами из тестов с высокими ставками из Нидерландов и Кении. Для оценки НМВП в тестах с высокими ставками могут использоваться как открытые, так и закрытые форматы заданий. Для всех форматов важно уделять внимание и время подготовке к разработке валидных ключей и дистракторов (закрытые задания) или валидных и функциональных схем оценивания (открытые задания).

Ключевые слова: НМВО, высокозначимое тестирование, валидность, надежность, таксономии, оценивание в контекстах.

References

1. OECD. *Programme for International Student Assessment (PISA)*. – 2024. - https://www.oecd.org/en/about/programmes/pisa.html
2. Biggs J. B., Collis K. F. *Evaluating the Quality of Learning: the SOLO taxonomy*. – New York : Academic Press. – 2014.
3. Case R. The Unfortunate Consequences of Bloom's Taxonomy // *Social Education*. – 2013. – Vol. 77 (4). – P. 196–200.
4. Soozandehfar S., Adeli M. A Critical Appraisal of Bloom's Taxonomy // *American Research Journal of English and Literature*. – 2016. – Vol. 2. – P. 1–9. – DOI:10.21694/2378-9026.16014
5. Tolman A. O., Kremling J. *Why students resist learning: A practical model for understanding and helping students*. – Stylus Publishing, 2017.
6. *Bloom's Taxonomy: Benefits and Limitations* // Intentional College Teaching. – 2021. – 30 April. - https://intentionalcollegeteaching.org/2021/04/30/blooms-taxonomy-benefits-and-limitations
7. Sultanova G. Assessment tasks in STEM education: integrating Bloom's Revised Taxonomy and Cognitive Load Theory // *Research in Science & Technology Education*. – 2025. - 10 Nov. – P. 1–18. – https://doi.org/10.1080/02635143.2025.2586564
8. Sweller J. Cognitive Load During Problem Solving: Effects on Learning // *Cognitive Science*. – 1988. – Vol. 12, № 2. – P. 257–285. – https://doi.org/10.1207/s15516709cog1202_4
9. Arwood E. L. *An Introduction to Pragmatic Assessment and Intervention for Higher Order Thinking and Better Literacy*. – UK : Jessica Kingsley Publishers, 2011. – 304 p.
10. Macat. – 2026. - https://www.macat.com/
11. Shin H. J., Li S., Ryoo J. H., von Davier A., Lubart T., Khalil S. The Nature and Measure of Critical Thinking: The PACIER Framework and Assessment // *Journal of Intelligence*. – 2025. – Vol. 13. – P. 1-16. – https://doi.org/10.3390/jintelligence13090113
12. OECD. *PISA 2015 Draft Mathematics Framework*. – March 2013. - https://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm

13. Leigh A. AI Question Generation: The Risks and Alternatives // *Grademaker*. – 2024. – 5 March. - https://www.grademaker.com/news/ai-question-generation-the-risks-and-alternatives/

14. Corbin T., Bearman M., Boud D., Dawson P. Assessment & Evaluation in Higher education: the wicked problem of AI and assessment // *Assessment & Evaluation in Higher Education*. – 2025. – P. 1–13. – https://doi.org/10.1080/02602938.2025.2553340

15. Kahn S. M., Hamer J., Almeida T. Generate: A NLG system for educational content creation // *Proceedings of the 14th International Conference on Educational Data Mining*. – 2021.

16. Black B., Bramley T., Suto I. The interrelations of features of questions, mark schemes and examinee responses and their impact upon marker agreement // *Assessment in Education: Principles, Policy and Practice*. – 2011. – Vol. 18 (3). – P. 295–318. – https://doi.org/10.1080/0969594X.2011.555328

17. Liu Q., Wald N., Daskon C., Harland T. Multiple-choice questions (MCQs) for higher-order cognition: Perspectives of university teachers // *Innovations in Education and Teaching International*. – 2024. – Vol. 61 (4). – P. 802–814. – https://doi.org/10.1080/14703297.2023.2222715

18. Bibler Zaidi N. L., Grob K. L., Monrad S. M., Kurtz J. B., Tai A., Ahmed A. Z., Gruppen L. D., Santen S. A. Pushing Critical Thinking Skills With Multiple-Choice Questions: Does Bloom's Taxonomy Work? // *Academic Medicine*. – 2018. – Vol. 93 (6). – P. 856–859. – https://doi.org/10.1097/ACM.0000000000002087

19. Stringer J. K., Santen S. A., Lee E., Rawls M., Bailey J., Richards A., Perera R. A., Biskobing D. Examining Bloom's Taxonomy in Multiple Choice Questions: Students' Approach to Questions // *Medical Science Educator*. – 2021. – Vol. 31, № 4. – P. 1311–1317. – https://doi.org/10.1007/s40670-021-01305-y.

20. Dieteren N., Jongkamp C. Higher-order thinking skills in high stakes testing: how to develop valid and reliable assessment instruments in 21st century contexts? // *42nd Annual Conference of the IAEA*. – Cape Town. – 2016.

Information about the author:
Nico Dieteren – Dutchtestologist, The Netherlands, dutchtestologist@outlook.com

Автор туралы мәлімет:
Нико Дитерен – Dutchtestologist, Нидерланды, dutchtestologist@outlook.com

***Сведения* об авторе:**
Нико Дитерен – Dutchtestologist, Нидерланды, dutchtestologist@outlook.com