Towards a Development of Automated Feedback and Diagnostic Assessment System in Support of Literacy Teaching and Learning

Girts Burgmanis[a], Dace Namsone[b] and Ilze France[c]
Interdisciplinary Centre for Educational Innovation, University of Latvia, Riga, Latvia

Keywords: Technology-Based Assessment, Diagnostic Assessment, Literacy, Feedback.

Abstract: Literacy knowledge and skills are crucial to students to understand complexities of the modern world. Furthermore, literacy can support and make more effective learning of discipline-specific knowledge. Literacy as a transversal skill makes it difficult for teachers to design assessment to gather information for feedback to facilitate literacy teaching and learning. In the last decade, technology-based assessment and automated text scoring has given the opportunity to develop diagnostic assessment systems providing immediate feedback enhancing students learning. This paper describes automated feedback and diagnostic assessment system in support of literacy teaching and learning. By validating three-dimensional framework of literacy assessment and evaluating measurement instrument quality, we provide evidence how advances in technology-based assessment and education studies can be implemented to design an online diagnostic assessment system providing immediate generic feedback to students and teachers.

1 INTRODUCTION

It is indisputable that in modern world the ability to deal with different forms of text and communicate meaning in written form or literacy is a crucial skill to successfully participate in the information society and improve academic achievement. Thus, it is important to support students who have trouble with literacy learning as soon as possible. Although it is a simple agenda, however to help students to recognize and improve errors or teachers to acquire information for interventions we need reliable and valid assessment and effective feedback.

Educational research shows that support for teaching and learning through diagnostic assessment (Walters, Silva & Nikolai, 2017) as well as immediate feedback (Hattie and Timperley, 2007) have very high potential. With development and application of technology-based assessment (TBA) and automated text scoring techniques in assessment, immediate feedback becomes more practical, less time consuming and human bias dependent.

In the present paper, we focus on theoretical foundation of automated feedback and diagnostic assessment system developed by Interdisciplinary Centre for Educational Innovation of University of Latvia. The system is designed for diagnostic purposes to assess students' literacy knowledge and skills in grade 4, 7 and 10 and provide immediate feedback on students' performance. In particular, we focus on validation of three-dimensional framework for literacy assessment and evaluation of a measurement instrument quality.

2 DEFINING LITERACY

Definition of literacy is vague and depends on the context. In scientific literature, there is a range of literacy definitions, from skills-based conceptions of functional literacy to broad definitions that integrate social, economic and political empowerment (Lonsdale & McCurry, 2004).

The most common understanding of literacy can be expressed as people's ability to read and write. This perspective on literacy is based on definition provided by the United Nations Educational, Scientific and Cultural Organisation (UNESCO) in

[a] https://orcid.org/0000-0001-5903-2283
[b] https://orcid.org/0000-0002-1472-446X
[c] https://orcid.org/0000-0002-3838-8157

Burgmanis, G., Namsone, D. and France, I.
Towards a Development of Automated Feedback and Diagnostic Assessment System in Support of Literacy Teaching and Learning.
DOI: 10.5220/0011858100003470
In *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023) - Volume 2*, pages 296-305
ISBN: 978-989-758-641-5; ISSN: 2184-5026

1950s, declaring that a person is functionally literate when he/she has acquired the knowledge and skills in reading and writing which enable them to engage effectively in all those activities in which literacy is normally assumed in his culture or group.

In the last two decades, through effort of the United Nations the meaning of literacy has changed. New definition of literacy proclaimed by 'United Nations Literacy Decade: education for all' (UN, 2002) emphasizes perspective that literacy as an individual skill is not only important for the increase of national productivity, but also for the development of sustainable, literate communities.

These changes in meaning of literacy reflect changes in understanding of what is text in digital era where traditional view of literacy as 'reading, writing, listening and speaking' is extended to viewing and representing (Ljungdahl & Prescott, 2009). Text in new literacy studies (e.g. Mills, 2010) is understood in a broader and more complex sense than previously and includes not only printed or digital written text, but also animations, movies, graphs, diagrams, images and maps. Thus, in modern society we expect that person with literacy skills will be proficient to read various modes of texts and their combinations, i.e., multimodal texts (Serafini, 2012), understand the message of text, reflect in oral or written form on text and its aims, and interpret various texts. This meaning of literacy is partly used by OECD PISA, defining reading literacy as: understanding, using, reflecting on and engaging with written texts, in order to achieve one's goals, develop one's knowledge and potential, and participate in society (OECD, 2019).

3 DIAGNOSTIC ASSESSMENT AND EFFECTIVE FEEDBACK

Historically, assessment has multiple purposes – including, to gather relevant information about student performance or progress, or to determine student interests to make judgments about their learning process. By using information from assessment, teachers can adjust their instructional practice according to student needs. Assessment for students can work also as motivational force to learn. Moreover, in recent years governments throughout the world use assessment data acquired through national or large-scale international testing (e.g. Trends in International Mathematics and Science Studies – TIMSS, Progress in International Reading Literacy Study – PIRLS, and Program for International Student Assessment – PISA) to evaluate standards of learning and to implement changes in national educational systems (Black & Williams, 2010).

Traditionally, to meet all previously mentioned purposes the summative assessment or assessment of learning (Hume & Coll, 2009) is used to acquire important data and information for decision making. Previous studies show that this approach has limited influence on learning experiences of students (Black & Williams, 1998) and if improvement of learning is paramount then the purpose of assessment should change from assessment of learning to assessment for learning which focuses on constructive feedback and on developing the student's capacity to self-assess and reflect on his/her own learning (Holmes-Smith, 2005; Ljungdahl & Prescott, 2009).

Usability of summative assessment for improvement of learning has several important deficiencies, including (1) long time between test administration and feedback, (2) limitations to use summative test results for intervention planning (Csapo & Molnar, 2019), (3) washback effect when teachers may teach directly for specific test preparation or learners adjust their learning strategies to the announced assessment (Leber et al., 2018), (4) harmful influence on school climate and teacher stress (Saeki et al., 2018).

Regular and immediate feedback acquired from test results of formative (Black & Williams, 2009) and diagnostic assessment (Ketterlin-Geller & Yovanoff, 2009) is most powerful tool to facilitate learning (Hattie and Timperley, 2007). Although formative assessment in educational studies is viewed as the most important source of information to support teaching and learning, however the difficulty with effective formative assessment is that it is very time consuming and rarely possible for busy classroom teachers to design (Treagust, 2006). Thus, diagnostic assessment most often is seen as a compromise between the need of an immediate feedback and an increased workload of teachers.

Distinction between formative and diagnostic assessment is vague. Most studies describe diagnostic assessment as assessment that focuses on problems, explores possible difficulties, assesses if students are prepared for a learning task, and measures prerequisite knowledge (Csapo & Molnar, 2019). Diagnostic assessment provides teachers with valuable information on what students know or do not know and for design of supplemental interventions for struggling students. At the same time, students receive feedback on their performance which provides opportunity to monitor their own learning (Walters, Silva & Nikolai, 2017).

Previous studies show that the effectiveness of feedback can be determined by several factors, such as the features of the feedback, the characteristics of the learners, and the nature of the tasks (Shute, 2008). Feedback features in previous studies address the timing of the feedback and the content of the feedback. In terms of timing, feedback can be immediate or delayed where immediate feedback is more effective than delayed (Kulik & Fletcher, 2016), because it can allow for a student to promptly recognize and understand conceptual and procedural errors (Dihoff, Brosovic & Epstein, 2003).

The content of the feedback can be two-fold and include: (1) simple information if answer is correct or incorrect, (2) additional information on student performance in test in overall or in single answer. Studies show that elaborated feedback is more effective and can significantly facilitate learning in classrooms (Maier et al., 2016, Lee et al., 2019). Moreover, the study of Zhu, Liu and Lee (2020) suggests that elaborated feedback is more effective for learning if it is adjusted to the context of each task (contextualized feedback). Several studies show that content of effective and useful feedback should include at least three components - indicate what student know or can do at the moment of the test (actual performance), what student need to know or do (desirable performance) and recommendation how to achieve desirable performance (Hattie and Timperley, 2007, Gibss & Simpson, 2004, Wiliam, 2006).

Technological transformation of education determines that technology-based assessment and, in particular, formative and diagnostic assessment is seen as a solution for support of everyday educational process and important source of information and feedback for students to monitor their progress.

4 TECHNOLOGY-BASED ASSESSMENT AND AUTOMATED SCORING

In the last two decades, the technology-based assessment (TBA) experienced rapid development. Education reforms all over the world implementing new approaches in learning, teaching and assessment (Cheng, 2020) along with the ICT development and rise of international comparative assessment (e.g. PISA, TIMSS etc.) significantly contributed to the transformation of TBA focus from examination of factual knowledge to 21st century skills.

The OECD PISA assessment had the greatest influence on the development of TBA by driving TBA development in participating countries and designing new methods and technologies in TBA. According to Molnar and Csapo (2019) history of TBA development had three stages. In the first stage, computer-based tests included digitized items, primarily prepared for paper-and-pencil assessment, as well as automated scoring and simple feedback. In the second stage, TBA included complex, real-life situations, authentic tasks, interactions, dynamism, virtual worlds, collaboration. Later stage of TBA development included search for solutions to elaborate support for personalized learning in the form of guidance and feedback to learners and teachers beyond test score.

Traditionally, technology-based assessment is perceived as a solution for providing immediate feedback to students. However, previous studies show that it is not exactly true, because the scoring of constructed response items is still complicated and time-consuming process even when using technologies (Gibbs & Simpson, 2005). In the last decade, the advances in machine learning allowed to develop and implement in technology-based assessment automated text scoring which is useful to handle several-sentence-long text responses.

Automated text scoring provides new possibilities for technology-based assessment including (1) to assess extended text responses in real-time, (2) reduce errors and biases introduced by human, (3) to save time, money and effort of teachers (Williamson, Xi, & Breyer, 2012, Liu et al., 2014). Automated text scoring works when an answer of student on constructed response item is assigned a score on a predetermined scale which can be established dichotomously (0 or 1) or in multiple scoring levels (0 to n) (Lee et al., 2019). After assigning a score to a response, student receives immediate feedback according to the scoring category framework or rubric. Automated text scoring mostly is used to assess the content of a response (Sukkarieh & Blackmore, 2009) or quality of writing (Bridgeman, Trapani, & Attali, 2012).

Latest findings from automated text scoring studies show that this approach is very effective to provide immediate formative feedback to students and support their scientific argumentation (Lee et al., 2019, Zhu, Liu & Lee, 2020).

Latest studies show that the best example of TBA system for supporting personalization of learning is online diagnostic assessment system – eDia developed by the Centre for Research on Learning and Instruction, University of Szeged (Csapo &

Molnar, 2019, Molnar & Csapo, 2019b). eDia system is designed to provide regular diagnostic information in three main domains of education - reading, mathematics, and science, from the beginning of schooling to the end of the 6 years of primary education. It is an integrated assessment system supporting assessment process, starting from item development, test administration, automated scoring, data analysis to feedback (Csapo & Molnar, 2019).

5 THREE-DIMENSIONAL FRAMEWORK FOR LITERACY ASSESSMENT

To provide credible feedback to support teaching and learning, it is important to design assessment framework which defines construct of concept what will be measured by the tests. In simple words, construct defines content of the assessment. Assessment framework is crucial to develop valid and reliable diagnostic measurement instrument for accumulation of evidence on student reading and writing skills and make meaningful indirect inferences on students' weaknesses and strengths.

To define assessment framework, we assume that literacy is complex concept, and its learning is multidimensional by nature. Moreover, according to the previous studies (Molnar & Csapo, 2019) learning and teaching in school have at least three clear goals – to develop cognitive abilities, to increase usability of knowledge and skills in various contexts and to transfer disciplinary knowledge important for students to navigate their social and personal lives as well as the world of ideas (Cuthbert, 2021).

Thus, we suggest that to support students to deal with textual information in different contexts, novel situations and to apply literacy knowledge and skills for problem solving, teachers have to assess and provide regular feedback on at least three dimensions of literacy learning. The first dimension, which should be assessed by diagnostic assessment, is disciplinary knowledge grounded in various contexts (science, history, literature etc.). The second dimension focuses on assessing students' skills to apply all types of textual information (texts, images, diagrams, maps etc.). To describe application of literacy skills more precisely we distinguished three categories (1) acquisition of textual information from various sources, (2) reasoning on textual information, (3) reflection on textual information and communication of textual information. The third dimension assess cognitive aspects of learning and represents students' progress in proficiency of literacy using SOLO (Structure of the Observed Learning Outcome) taxonomy (Biggs & Collis, 1982).SOLO taxonomy is directly grounded in constructivism theory proposing that learning grows cumulatively in stages in which the learned content is increasingly complex (Leung, 2000). SOLO taxonomy is useful to identify which stage student attained and describe two aspects of learning – quantitative or how much details students know, and qualitative or how well students put together that detail. SOLO taxonomy include five levels:

- Pre-Structural – The student's response include bits of unconnected information and represents that he/she does not have any kind of understanding on topic.
- Uni-Structural - The student's response include one relevant aspect of the subject or task and represents that he/she can make simple and obvious connections.
- Multi-Structural - The student's response include evidence that he/she can understand several separate aspects of the subject or task.
- Relational - The student's response include evidence that he/she can make relations between several aspects and demonstrates an understanding of the topic.
- Extended Abstract - The student's response demonstrate that he/she can make connections not only within the given subject field, but also make connections beyond it and transfer ideas to new areas.

Rationale to adopt SOLO taxonomy to develop third dimension of framework over Bloom's taxonomy were based on focus of assessment system. Bloom's taxonomy is useful to develop assessment and categorize items by what student must do in order to demonstrate learning, however, SOLO taxonomy enable not only to discriminate items, but also students' responses on items. For example, the same item by several students can be answered in different levels and demonstrate different understanding on subject. In this case Bloom's taxonomy can be used to determine overall complexity of item, but SOLO taxonomy to identify what students can do or understand and what they cannot. Thus, the use of SOLO taxonomy is crucial for diagnostic purposes, works as structured approach to evaluate constructed response items, and is useful for developing rubric for automated text scoring.

Third dimension of framework is useful for further development of automated feedback and scoring module, because it enables not only to create test items with various ranges of difficulty (four SOLO levels), but also to create detailed description

of students' progress in proficiency of literacy knowledge and skills.

6 AUTOMATED FEEDBACK AND DIAGNOSTIC ASSESSMENT SYSTEM

The system was built in 2021 in collaboration between the Interdisciplinary Centre for Educational Innovation of University of Latvia and business enterprise 'Izglītības sistēmas', owner of the digital school management system 'e-klase.lv' in Latvia.

Automated feedback and diagnostic assessment system is a technology-based, learning-centred and integrated assessment system consisting of three modules: (1) test editing module, (2) online test delivery module, (3) scoring module, (4) feedback module.

The system is designed for diagnostic purposes to assess students' literacy knowledge and skills in grade 4, 7 and 10. Currently, the development of system reached the second stage of testing when the feedback module will be tested. In the first stage of testing, we tested framework of literacy assessment, item writing and test editing module, online test delivery module and scoring module.

The items were written and saved in open source software GeoGebra applet and linked with system's test editing module using applet ID generated by GeoGebra. The rationale for the choice of GeoGebra as item writing module was to have opportunity to develop system further for the use of numeracy diagnostic assessment in the future. According to assessment framework, each item has three attributes - proficiency level, category of application of literacy skills and context (history, literature, science etc.).

The system contains test editing module that is intended to verify items for further use in online test delivery module, as well as for selection of items from item bank and construction of diagnostic tests. Items typically are selected by test developer (e.g. researcher or expert) based on attributes describing item proficiency level and category of application of literacy skills to design test on particular context.

Students complete each diagnostic test in school ICT classrooms. The tests can be ran by students using computers equipped with an internet browser, keyboard, mouse and screen. To access test, students need to login in the system by using their 'e-klase.lv' user login details. Each test can be administered by teachers who can choose the day and time when students have access to the test and complete it.

The system is designed for both automated and human scoring. System included two types of automated scoring with purpose to provide immediate feedback. First type of automated scoring was more traditional and designed to score multiple-choice items. Automatic text scoring was designed to score constructed response items. Students' constructed responses submitted online were randomly selected by the system for expert scoring and then processed by automatic text classifier trained using supervised machine learning and automated model-building processes.

Based on the autoscores, the students can receive two types of feedback: (1) on the proficiency level of literacy learning in overall, (2) on the proficiency level of each literacy knowledge and skill application category. Feedback consists of three descriptors describing what student can do (actual proficiency level), what he/she cannot do (next proficiency level) and a recommendation for improvement how to reach next proficiency level. An IRT model (Rasch model) is used to establish assessment scales and five proficiency levels – not proficient, below proficient, approaching proficient, proficient and exceed proficient.

7 METHODS

7.1 Participants

The sample of study included 190 students from grade 7. These students where from four Latvian secondary schools. To ensure representative study sample, all schools were selected based on longitudinal performance in national level assessment in Latvian language. All participants completed diagnostic assessment consisting of two tests examining literacy learning in science and history contexts. The data collection was done by using online assessment system.

7.2 Procedure

At the beginning of the assessment, students were provided with instructions about the usage of the system, and they were allowed to try it by completing three test items. The assessment took place in the schools' ICT labs using the available school infrastructure. Testing sessions were supervised by teachers. Teachers could choose how to administer tests of assessment, i.e., one test per day or both tests in one day. When the students had completed each test, they submitted their answers for automated

scoring. All submitted answers were autoscored. Automated text scoring was possible, because the automated text rater for each of 14 constructed response items was trained prior to implementation of the items in online test. To train automated text rater, we collected 317 students' responses on all 14 constructed response items prior to this study using paper-and-pencil version of items. These responses were scored by two experts for each test, based on rubrics developed according to the tree-dimensional literacy assessment framework.

7.3 Instruments

Each test was developed to be completed in 45 minutes. Science test consisted of 8 items and history test of 15 items. All test items were developed according to the three-dimensional framework. Thus, each item represented attributes of all three dimensions of literacy learning (see Table 1).

Table 1: Structure of the diagnostic assessment.

Domains of literacy learning	SOLO I	SOLO II	SOLO III and IV
Acquiring textual information	S1, H1, H6, H8	S2, H3, H10, H11	S7
Analytical reasoning	S3, H9	H2, H4, H7	S4, H12
Communication of information	H5	S5, H13, H14	S6, S8, H15

Note: Character (context, i.e., disciplinary knowledge): S – Science; H – History; Number describes item number in test.

All tests included both multiple choice items and constructed response items (see Figure 1).

Figure 1: An example (history test) of multiple-choice question intended to measure acquiring of textual information (item 1.1. in figure) in SOLO I level and analytical reasoning (item 1.2. in figure) in SOLO II level.

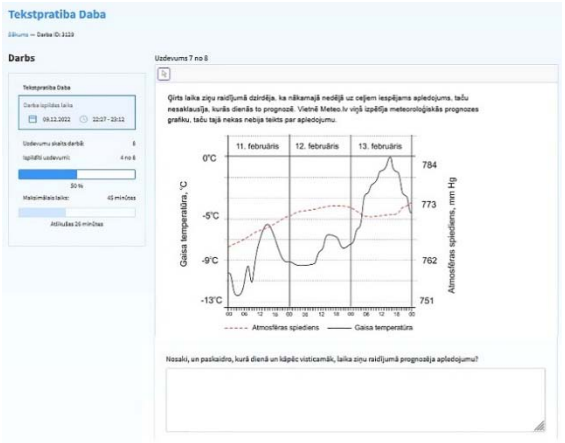

Figure 2: An example (science test) of constructed response intended to measure Communication of information in SOLO III level.

7.4 Analysis

In this study, to establish evidence on construct validity, i.e. correlation between dimensions of three-dimensional framework, the bivariate correlation was run. To examine the set of items included in assessment to measure numeracy learning, the Item Response Theory (IRT) and Rasch analysis for the Partial Credit Model (Masters, 1982) were used. For the statistical analysis the IBM SPSS software version 26.0 and WINSTEPS version 4.6.2 (Linacre, 2015) was used.

8 RESULTS

8.1 Validation of Three-Dimensional Framework for Literacy Assessment

The validity of three-dimensional framework or construct for literacy assessment was established by bivariate correlations (see Table 2). From Table 2 it can be seen that correlations between dimensions and their domains are medium to high ranging from .40 to .89.

Results show that correlations between disciplinary knowledge domains (r = .42), proficiency levels of literacy knowledge (r = .50 - .56) and application of literacy knowledge (r = .46 - .51) domains are medium.

Strongest correlations between disciplinary knowledge domain and its proficiency levels representing cognitive dimension are for history (r = .72 - .89) and medium for science (r = .45 - .73).

Similar correlations can be seen also between disciplinary knowledge and application of literacy knowledge dimensions. Stronger relationship can be seen for history (r = .77 - .79) than for science (r = .40 - .67) domain.

Relations between domains of proficiency levels of literacy knowledge and application of literacy knowledge range from medium to highly medium (r = .49 - .77). Thus, we can conclude that the correlation values in Table 2 contribute to the evidentiary support for the convergence of scores representing related constructs.

Table 2: Relations between results in the three dimensions of literacy knowledge learning.

Domains	H	S1	S2	S3	AI	AR	CI
S	.42	.50	.45	.73	.40	.52	.67
H		.73	.89	.72	.79	.77	.77
S1			.56	.50	.68	.65	.55
S2				.53	.76	.64	.74
S3					.49	.72	.77
AI						.51	.46
MR							.48

Note: All correlation coefficients are significant at level p<0.001. Disciplinary knowledge - H: history, S: science; Application of literacy knowledge - AI: acquiring textual information, AR: analytical reasoning, CI: communication of information; Proficiency levels of literacy knowledge - S1: SOLO 1, S2: SOLO 2, S3/4: SOLO 3 and SOLO4.

8.2 Evaluation of a Measurement Instrument Quality

Rasch analysis was used to evaluate quality of measurement instrument. To evaluate how well the test items are defining literacy, we used Wright maps which show difficulty of items and student's performance on the same linear scale.

Figure 3 show that students and items are located at the same level of the scale, i.e., ability level of students is equal to difficulty level of item. It means that students have 50% chance to correctly answer the item and diagnostic assessment discriminate low performing and high performing students very well. Wright map shows that diagnostic assessment consisting of two tests include items representing all difficulty levels or literacy learning proficiency levels. Two improvements which can make measurement better and close the difficulty gaps on the line are to include (1) two very difficult items between items H5 and S7, (2) at least one very easy item between items S2 and H9.

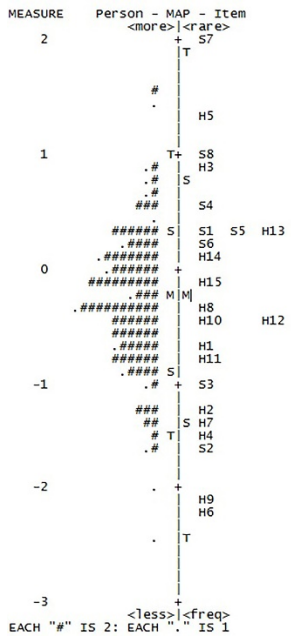

Figure 3: The Wright map of the diagnostic assessment items used in the present study.

Another approach to evaluate measurement instrument quality is to examine how items fit to the Rasch model (Boone, 2016). According to Boone (2016), fit analysis determines if difficult items are harder to answer than easy ones regardless of student's ability levels, and if items do not fit the model, we can say that they measure more than one variable.

To evaluate compliance of item to the Rasch model, most commonly fit statistics (e.g., MNSQ Item Outfit, MNSQ Item Infit) is used. Although various studies share similar opinion and suggest that evaluation of Outfit MNSQ statistic for each item is guiding principle, however the benchmark to say that the item misfit the model and is not useful for measuring variable varies between 1.3 (Boone, 2016) and 1.4 (Bond & Fox, 2013).

Table 3 shows that two items (S7 and H3) manifest excessively high levels of outfit and can be considered to be removed. After additional expert analysis of items, we concluded that both items measure not only literacy learning in context of science and history, but also require from students to demonstrate deep science and history knowledge and skills. It means that both items measure more than one variable at time. This assumption is also approved by Point-measure correlations which show how well items measure construct. Both items show negative Point-measure correlation values. It means that these items primarily do not measure literacy learning.

Table 3: Rasch analysis: infit, outfit statistics and point measure correlation for each diagnostic assessment item.

Item	Measure	Infit (MNSQ)	Outfit (MNSQ)	PMC
S1	.58	1.10	0.97	.29
S2	-1.31	.97	.89	.32
S3	-.71	.91	.89	.41
S4	.87	1.08	1.04	.43
S5	.65	.65	.69	.33
S6	.51	.94	.93	.54
S7	2.30	1.07	1.66	-.07
S8	1.32	.96	1.02	.28
H1	-.42	.99	.98	.29
H2	-1.01	1.01	.99	.24
H3	1.23	1.14	1.34	-.03
H4	-1.23	.97	.92	.31
H5	1.60	.92	.87	.44
H6	-1.90	.98	.95	.24
H7	-1.10	.99	.98	.26
H8	-.08	1.09	1.09	.13
H9	-1.71	1.03	1.08	.16
H10	-.13	1.14	1.12	.39
H11	-.47	1.21	1.21	.43
H12	-.16	1.20	1.17	.46
H13	.58	.72	.73	.36
H14	.40	.85	.93	.48
H15	.18	1.07	1.07	.53

Note: PMC – Point Measure Correlation

9 CONCLUSION

Considering rapid development of TBA and demand from education sector, where, at the beginning of 21st century, the learning paradigm has changed to student-centred learning and learning of transversal skills, we need to make teaching, learning and assessment more effective and personalized.

To support literacy teaching and learning, we developed an automated text scoring enabled diagnostic assessment system which can provide feedback to students, as well as to teachers immediately after completing a test. Our system at this stage provide literacy diagnostic assessment consisting of two tests (history and science) to assess students' literacy learning.

In this paper, we provide evidence that literacy learning can be validly described in three-dimensions of learning: disciplinary knowledge, application of literacy knowledge and skills, and proficiency level, which all are related constructs. Rasch analysis confirmed that literacy assessment consisting of two tests is appropriate to measure students' literacy learning. Developing items for literacy assessment in future, it is important to verify that items measure primarily literacy learning, not disciplinary knowledge.

Finally, we can conclude that system can be used to support students' personalized learning and teachers' instructional decisions in literacy teaching and learning. However, we should admit that automated feedback and diagnostic assessment system developed by experts and researchers could create several challenges for teachers to use it for intervention planning. Teachers are not always adequately prepared and capable to use student achievement data in meaningful way to improve student learning in the classroom and improve overall student achievement (Dunlap & Piro, 2016, Lockton, Weddle & Datnow, 2020). This is not functional weakness of system, but may have significant impact on implementation of system. Thus, in implementation phase we have to plan and provide professional development programmes to teachers to increase their (1) data and assessment literacy; (2) familiarity with system.

Furthermore, at this phase of system testing show that usability of system and reliability of data provided by system directly depends not only on students' digital skills in general, but also on their familiarity with GeoGebra environment integrated in online test delivery module. GeoGebra environment for younger students can be challenging and create situation that diagnostic assessment could measure domains out of the scope of three-dimensional framework (e.g., digital skills, ingenuity). In this study we familiarize participants with system offering to them complete three test items before they performed the main test. However, this solution have to be studied further to understand – does completion of few training items in training environment of system prior diagnostic test could be enough for students to demonstrate their best performance including younger students (e.g. grade 4) and to introduce them with the GeoGebra environment.In the future, we plan to design and carry out several new studies where literacy diagnostic assessment will be supplemented with tests from other disciplines to provide more detailed view on students' literacy learning. Moreover, we plan to carry out study where the effect of feedback on literacy learning provided by system will be empirically tested.

ACKNOWLEDGEMENTS

The research was supported by the European Regional Development Fund's project 'IT-based support system prototype for providing feedback and improve student performance in literacy and numeracy acquisition', Project No. 1.1.1.1/19/A/076.

REFERENCES

Biggs, J., and Collis, K. (1982). *Evaluating the Quality of Learning: the SOLO taxonomy*. New York: Academic Press.

Black, P., and Wiliam, D. (2009). Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability, 21(1), 5–31.

Bond, T. G. and Fox, C. M. (2013). *Applying the Rasch model: Fundamental measurement in the human sciences*. New Jersey: Lawrence Erlbaum Associates Publishers, 2013.

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how?. *CBE—Life Sciences Education*, 15(4), 1.-7.

Bridgeman, B., Trapani, C., and Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, 25(1), 27-40.

Cheng, Y. C. (2020). Education reform phenomenon: A typology of multiple dilemmas. In Fan, G., and T. Popkewitz (Eds.), *Handbook of education policy studies* (pp. 85-109). Singapore: Springer

Csapó, B., and Molnár, G. (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia system. *Frontiers in psychology*, 10, 1522.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., and Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st Century skills* (pp. 143–230). New York, NY: Springer.

Cuthbert, A. S. (2021). Disciplinary knowledge and its role in the school curriculum. In A. S. Cuthbert and A. Standish, *What should schools teach?* (pp. 15-37). London: UCL Press.

Dihoff, R. E., Brosvic, G. M., and Epstein, M. L. (2003). The role of feedback during academic testing: The delay retention effect revisited. *The Psychological Record*, 53(4), 533-548.

Dunlap, K., Piro, J. (2016). Diving into data: Developing the capacity for data literacy in teacher education. *Cogent Education*, 3(1), 1132526.

Gibbs, G., and Simpson, C. (2005). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, (1), 3-31.

Hattie, J., and Timperley, H. (2007). The power of feedback. *Review of educational research*, 77(1), 81-112.

Holmes-Smith, P. (2005). *Assessment for Learning: Using statewide literacy and numeracy tests as diagnostic tools*. Paper presented at the Using data to support learning research conference, Mlbourne.

Hume, A. and Coll, R. K. (2009). Assessment of learning, for learning, and as learning: New Zealand case studies. *Assessment in Education: Principles, Policy & Practice*, 16(3), 269-290.

Ketterlin-Geller, L. R., and Yovanoff, P. (2009). Diagnostic assessments in mathematics to support instructional decision making. *Practical Assessment, Research, and Evaluation*, 14(1), 1-11.

Kulik, J. A., and Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Review of educational research*, 86(1), 42-78.

Leber, J., Renkl, A., Nückles, M., and Waschle, K. (2018). When the type of assessment counteracts teaching for understanding. *Learning: Research and Practice*, 4(2), 161-179.

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real‐time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, 103(3), 590-622.

Leung, C. F. (2000). Assessment for learning: Using SOLO taxonomy to measure design performance of design & technology students. *International Journal of Technology and Design Education*, 10(2), 149-161.

Linacre, J., M. (2015). *Winsteps Rasch Measurement (computer software)*, Beaverton, Oregon: Winsteps.com

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., and Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice*, 33(2), 19-28.

Ljungdahl, L., and Prescott, A. (2009). Teachers' use of diagnostic testing to enhance students' literacy and numeracy learning. *International Journal of Learning*, 16, 461-476.

Lockton, M., Weddle, H., Datnow, A. (2020). When data don't drive: teacher agency in data use efforts in low-performing schools. *School Effectiveness and School Improvement*, 31(2), 243-265.

Lonsdale, M., and McCurry, D. (2004). *Literacy in the new millennium*. Adelaide: NCVER.

Maier, U., Wolf, N., and Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*, 95, 85-98.

Masters, G., N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

Mills, K. A. (2010). A review of the "digital turn" in the new literacy studies. *Review of educational research*, 80(2), 246-271.

Molnar, G. nad Csapo, B. (2019). Making the psychological dimension of learning visible: using technology-based assessment to monitor students' cognitive development. *Frontiers in Psychology*, 10, 1368

Molnár, G., and Csapó, B. (2019b). Technology-based diagnostic assessments for identifying early mathematical learning difficulties. In Fritz A., Haase V., and Rasanen P. (Eds.), *International handbook of mathematical learning difficulties* (pp. 683-707). Cham: Springer.

OECD, (2019). *PISA 2018 Assessment and Analytical Framework*, Paris: OECD Publishing.

Saeki, E., Segool, N., Pendergast, L., and von der Embse, N. (2018). The influence of test-based accountability policies on early elementary teachers: school climate, environmental stress, and teacher stress. *Psychology in the Schools*, 55, 391–403.

Serafini, F. (2011). Expanding perspectives for comprehending visual images in multimodal texts. *Journal of Adolescent & Adult Literacy*, 54(5), 342-350.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, 78, 153–189.

Sukkarieh, J. Z., & Bolge, E. (2010). Building a textual entailment suite for evaluating content scoring technologies. *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (pp. 3149–3156). Paris: European Language Resources Association.

Treagust, D. F. (2006). Diagnostic Assessment in Science as a Means to Improving Teaching, Learning and Retention. UniServe Science–Symposium Proceedings: Assessment in science teaching and learning (pp. 1–9). Sydney: Uniserve Science, Sydney.

Walters, S. R., Silva, P., and Nikolai, J. (2017). Teaching, learning, and assessment: Insights into students' motivation to learn. *The Qualitative Report*, 22(4), 1151-1168.

Wiliam, D. (2006). Assessment: Learning communities can use it to engineer a bridge connecting teaching and learning. *The Learning Professional*, 27(1), 16-20.

Williamson, D. M., Xi, X., and Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: issues and practice*, 31(1), 2-13.