Evaluating Educational Standards using Assessment “with” and “through" Technology

Lena Frenken[1], Paul Libbrecht[2], Gilbert Greefrath[1], Daniel Schiffner[3], Carola Schnitzler[4]

[1]University of Muenster, Institute for Mathematics and Computer Science Education, l.frenken@uni-muenster.de; [2]IUBH Fernstudium; [3]DIPF | Leibniz Institute for Educational Research and Information; [4]IQB Institute Educational Quality Improvement, Germany

This paper reports on a feasibility study of creating a standardised assessment instrument to evaluate students' competencies found in the German national standards. The study aimed at combining widespread tools in math-classes, such as dynamic geometry and spreadsheets, in an integrated and computer-driven way. We report on the mathematical and technical feasibility: What limits were reached, and which opportunities have appeared? The report provides indications that a development process is feasible but that an attention to the task description is required, as the student may be unaware of the manipulations to perform tasks.

Keywords: assessment, educational standards, affordances, dynamic geometry, spreadsheets

Introduction

The possibilities of using digital media in (mathematical) learning processes increases more and more. Therefore, large-scale assessments have to be adapted to new ways of learning and to new competencies. The current technical state of the art is to focus on accepting a final answer from students (Pelkola, Rasila & Sangwin, 2017) though the collection of log-data is considered to be enhancing assessment.

In Germany a large-scale assessment called VERA is conducted based on a paper-pencil-test once a year in grade 8 (14 y old) (IQB, n.d.). Turning it into a digital test environment could bring many advantages: assessing mathematical digital competency (Csapo, Molnar & Toth, 2009; Geraniou & Jankvist, 2019), using rich and dynamic items, or the integration of automatic scoring (Drijvers, 2018). This paper is based on a first feasibility study with an evolved test instrument that combines key features of the existing paper-pencil-assessment with innovative ideas.

BACKGROUND

Standardized competency assessment is a form of testing that aims to measure the competencies reached by all testees in a comparable way. The results can inform on the attainment of teaching (as is the case of TIMSS or PISA studies, but also as any examination), on the general competency considered important (as is the case for the PIAAC study) or for other research purposes. Standardized assessment has often been made with paper and pencil and this remains the dominant practice for mathematics competency testing as the manipulation of mathematical objects on computers remains fragmented and isolated. Nevertheless, both theoretical considerations as

well as studies demonstrate the strong potential of testing with a computer. This applies particularly to the higher objectivity of automatic scoring and the larger amount of information obtained through log-files (e.g. as described by Goldhammer & Zehner, 2017). Aspects such as the individual task solving behavior, the testee's self-confidence or the dependency on a particular expression media can be assessed. Among other aspects, this feasibility study aims at exploring how achievable it can be to obtain such extra information.

In mathematics education, assessment constitutes a large area. The notion of computer-based mathematical assessment becomes more commonplace by the fact that computers are often used to support mathematical learning activities. Stacey and Wiliam (2012) differentiate *assessment with digital technology*, where the computer-tools (even calculators) support mathematical processes but not assessment processes, from *assessment through digital technology*, where it is driven by computer activity.

Sangwin et al. (2009) propose ways to automatically evaluate certain answer types and create a competency model based on the use of particular assessment tools. Such methods are mostly applied with formative assessment in which results help to enhance the learner's competencies rapidly. In contrast, little best practice is known for summative evaluations of mathematical competencies, whereby the technical state of the art is to focus on accepting a final answer from students (Pelkola, Rasila & Sangwin, 2017). Fine-grained analysis on the manipulation of mathematical tools is an evolving domain: Multiple research around intelligent tutoring systems have given the rise to successful training tools such as the Algebra Tutor (Koedinger et al., 2008). They are largely specialized and almost impossible to integrate within a summative evaluation where the overarching goal is to evaluate the breadth of competencies and not to depend on specialized tools, each including specialized user-interfaces. Since this study aims at a broad competency evaluation, a versatile and generic tool to perform standardized assessment has been chosen: the product CBA-ItemBuilder (Rölke et al., 2012).

However, as noted by Drijvers et al. (2016), widespread digital tools for performing mathematics exist and are even part of learners' everyday lives. The ICILS studies showed that German schools are equipped below-average regarding to technology-related resources for both teaching and learning (Fraillon et al., 2019). The strived body of competencies, the national educational standards (NES), is based on a subject-related normative competence structure model that explains which competencies students should gain until the finalization of different grades. The NES are oriented at general educational aims. In the case of mathematics education, the NES describe a competence model with three dimensions: competencies, basic concepts (i.e., contents) and levels of requirements (i.e., difficulties) as shown in Figure 1 (KMK, 2003). Aiming at evaluating the educational system and investigating which competencies students have achieved by certain grade levels, the nationwide paper-pencil test *VERA* (VERgleichsArbeiten [comparison test]) is carried out in Germany. VERA is based on the described normative model of

competence. Though the subject-related use of digital tools and media is not mentioned directly in this model, the basic concepts can be divided into different facets where it is mentioned partly. Nevertheless, more and more teachers integrate digital media to enhance learning processes. Therefore, the integration of tools very similar to the learned tools appears to be a good approach to the paradigm of assessment with technologies while gaining the benefits of assessment through digital technology: Using technology with affordances (in the sense of Kaptelinin, 2013) that are well-known to learners such as radio buttons, or that are either part of familiar digital tools for doing mathematics or very similar to them. In the case of the NES, this includes calculators, spreadsheets and dynamic geometry systems.

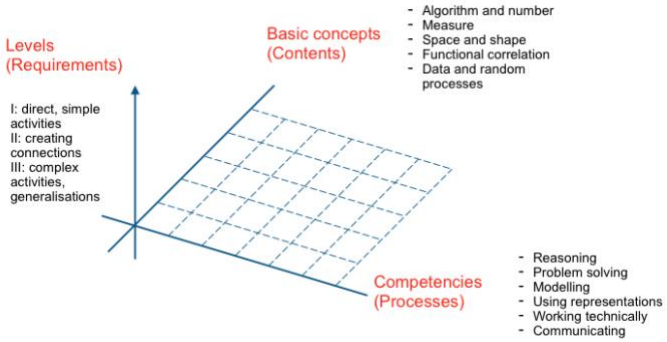

Figure 1 The normative competency structure model of the NES for the subject mathematics (see KMK, 2003).

The analysis of student answers using these technologies is only partially widespread. While the use of dynamic geometry systems for learning is common, the analysis of the correctness is not: Pioneering works such as in the ThEdu workshops, of Kovács, Recio and Vélez (2018), or of Kortenkamp and Richter-Gebert (2004) have not yet yielded a widespread applicability (Pelkola, Rasila & Sangwin, 2017) even though they demonstrate elementary validation strategies such as the use of simple predicates to indicate when two points are close to each other. Finally, there seems to be a need of analyzing the use of spreadsheets for learning purposes.

Our feasibility study explores the creation of a standardized assessment tool that can be deployed under general school conditions to assess mathematical competencies, including the application of digital tools and, therefore, even assessing mathematical digital competency (Geraniou & Jankvist, 2019), by means of a computer-based tool. Since we aim at using the schools' infrastructure, this exploration implies certain technical challenges: Some parts of the infrastructure may break because of the computer resources used by several actors (e.g. the bandwidth consumption, but also the installation of incompatible software). Moreover, the intent to combine *assessment with technology* with *assessment through technology* raises several validity concerns (Drijvers, 2018). In contrast, the combination also offers the opportunity to reach educational validity compared to using technically simple MC-items only (Sangwin & Jones, 2017).

RESEARCH DESIGN

Taking into account the described opportunities and limitations, we have set forth in this project to investigate the feasibility of assessing the dimensions of the NES using digital means despite the under-average equipment. The special demands of this as well as the associated paper-pencil test instruments played a special role in

developing the test. They had an impact on the item design process, the technical realization and its technical deliveries. Besides exploring which contents of the NES can be assessed at all, we intended to investigate technical opportunities of embedding dynamic geometry and spreadsheet software. Because of its many opportunities, automatic scoring and a log-data-collection empowered by computer-based-assessment were included. The following question defines the focus of the first feasibility study: To what extent can competencies of the NES be assessed by digital means and integrating a variety of digital tools that are common to the testees?

To answer this question by means of an explorative study, we developed a construction process spanning from the design of items to the delivery. During the process we constantly investigated technical possibilities and limits. The first step was to analyze the contents of the NES that could be assessed at all or could be assessed better than with paper-pencil. In contrast to comparable studies (see Csapo, Molnar & Toth, 2009; Pelkola, Rasila & Sangwin, 2018), the focus was not laid on designing items that can be assessed in a paper-pencil test as well or using items that already exist. Instead, innovative items with integrated digital tools or enhanced tasks through embedding video or audio material should be designed. Concluding, this study aimed at increasing the sophistication: dimension F should be reached in the sense of the assessment possibilities proposed by Hoogland and Tout (2018). On this basis, the item authors, who most work as mathematics teachers in Germany, have designed tasks that – in terms of structure – are mainly inspired by the paper-based VERA tasks, but which incorporate the innovative possibilities mentioned above. The item authors did not use tools of constructing tasks completely in a digital format but designed documents with descriptions of the items and prepared files for the digital tools. Once the documents were developed, two rounds of reviews were conducted by experts in mathematical didactics. Between the two rounds a detailed discussion on the items took place and the items were revised. This process of item construction followed partly the established development process of paper-based items (see Rupp & Vock, 2007). It was repeated three times and the subsequent selection followed a few criteria: Firstly, diverse media with a broad spectrum (dynamic geometry, spreadsheets, video, audio, picture, calculator) should be included so that technical possibilities and limits as well as students' usage of different embedded media can be evaluated; secondly, different facets of the NES (different contents, competencies and difficulties) should be assessed.

Following the motivation of assessing with technology (Drijvers et al., 2016), several digital tools for mathematics have been considered. One of each category was expected. Because the brief testing time, test-specific learning should be limited to a minimum. It was, thus, important that the tool affordances (see Kaptelinin, 2013) are similar to those of the tools known by the testees. Moreover, to avoid the need for a technique-oriented workflow, the tools were to be integrated in a webpage.

For dynamic geometry, the strong diffusion of GeoGebra at schools and its abilities to be used on the web made it a de facto candidate. It is important to note that a

GeoGebra construction embedded in a webpage, just as an activity that can be found among the GeoGebraTube server, is not necessarily including all functions of a desktop application. Multiple parameters allowed to restrict the set of actions.

For spreadsheets, the dominant tool is Microsoft Excel, a tool that only lives in its entire function set on desktops of Windows and macOS computers. Desktop alternatives often used at school include OpenOffice and LibreOffice. While multiple web-based spreadsheet services exist such as Collabora or Office365, we have either evaluated their incompleteness, bandwidth demand, license or incompatibility with regards to privacy. Only two tools remained with an open-source license and with an almost complete runnability on the client: EtherCalc or OnlyOffice. The latter was chosen for its greater visual and functional similarity to the dominant tool.

Using the item authoring tool CBA-ItemBuilder the chosen tasks were converted from design sketches into the digital format by item-implementors embedding all the external resources. This was followed by an internal reviewing process focusing on the technical problems: loops of revisions followed, so that a test instrument was designed for a 45-minute test period and could be conducted in nine classes. In total, 229 students took part. Beforehand, a system check was able to show that the schools were suitable for participation by checking technical requirements and carrying out test-like scenarios on randomly selected computers within the participating schools. Both system check and testing were observed and documented by the test leaders.

Automatic scoring was applied (with MCQs, with GeoGebra predicates). The logfiles generated by the CBA-ItemBuilder were converted into Excel-files and GeoGebra- and OnlyOffice-snapshots were extracted to visualize the final stage.

RESULTS

On the basis of the described research process, especially two strands of first results can be presented: concerning the assessment and the technical realization.

Assessment: First of all, it can be stated that this study was able to assess elements from all the basic concepts (see Figure 1): carry out targeted measurements in their environment, take measurements from source material, use them to carry out calculations and evaluate the results and the methods in relation to the situation; operate mentally with lines, surfaces and solids; draw and construct geometric figures using appropriate tools such as compass, ruler, triangle ruler or dynamic geometry software; use percentage calculation for growth processes (for example, interest calculation), also using a spreadsheet; systematically collect data, record it in tables and present it graphically, also using appropriate tools such as software. Those contents were spread over the different competencies and levels of requirements.

It could be observed that students have been comfortable using the web-browsers and the standard input affordances such as plain text fields or radio buttons. However, several students were unclear how to enter simple mathematical formulae such as the multiplication sign in a regular text field with the keyboard (normally written with the

special characters · or ×). We assume that providing a symbol bar may be enough but the need to input more complex formulae (e.g. roots or fractions) might appear. Based on the analysis of GeoGebra-snapshots, difficulties can be extracted: In cases where a construction was presented without specific buttons (e.g. a cube in a 3D space), some learners have simply not discovered the possibilities to explore a solid from all its facets. In others, testees moved provided elements such as points without any discernable mathematical activity. In cases where a construction was required to be done either by using the adequate tool directly or by constructing the different steps (a perpendicular bisector), some testees simply moved provided elements or did something not appropriate for the task. However, the assessment seems to have been successful with dynamic geometry constructions when the operations were simple and explained by small sentences (e.g. "drag the dots according to the cube" or "move the girl to estimate the height"). Therefore, the difficulties may sometimes be attributed to a lack of appropriate usage of dynamic geometry tools, but sometimes also to missing (not digital) mathematical competencies or mathematical knowledge. As for the spreadsheet, even simple tasks such as entering `=4+5` could not be executed correctly; instead the testees entered `4+5` or `4+5=` which lead them to ask the test leader why the calculation was not executed.

Technical Realization: Among the biggest technical challenges was the use of the schools' infrastructure because of the reported equipment limitations in Germany.

Thanks to the new runtime technology for the CBA-ItemBuilder concluding its development (using contemporary frameworks for JavaScript), the delivery technology could be refined with lower dependency on the network: Ensuring that most web-resources are stored in cache prior to starting and delivering the measured assessment data (results and logs) in an asynchronous way. This study has provided good signs that this approach is doable in schools: In classes where preparation in advance was done only a little significant lag was perceived. In a few cases, web-browsers became unstable; changing the computer was then the go to solution.

The web-embedding nature has shown to be viable. In this study, it was based on the principle of iframes (webpages in webpages) which communicate to the CBA-ItemBuilder to send and receive their data. As long as an introspection of the tools' state was technically feasible, it was possible to gather the changes of interesting objects (e.g. movements of points); this was the case for GeoGebra but not yet for OnlyOffice. However, the storage of state was possible in both cases, thus, enables evaluators to view the last created state (a geometry construction or spreadsheet).

CONCLUSION AND FUTURE WORKS

Concluding the feasibility study on evaluating the NES, a first analysis showed that an assessment of several facets seems doable as a large-scale assessment in future. Though technical realizations (e.g. the input of formulae) have to be developed further and affordances should be adapted to student habits, first the Item Response Theory is used to investigate the scalability. This would be a step in ensuring the

quality criterions and therefore making the tests comparable (Drijvers, 2018). In order to estimate why students were not familiar with all integrated tools, it is planned to conduct a survey on the use of digital media and tools in mathematics education in connection with an upcoming study. Further, for this upcoming study a greater amount of participating schools (20) is planned. Moreover, the workflow described above is going to be adapted, so that the item authors directly design tasks using the CBA-ItemBuilder and embedding the digital tools. We expect to make the workflow more effective and problems regarding the affordances or the technical realization faster to detect and handle in this way; this enriches the ongoing ItemBank conceptual development (Chituc et al., 2019), but requires more item-authoring capabilities. In summary, the goal to assess the NES for mathematics education with and through technology requires further development on technical aspects as well as on considerations and studies about how to ensure the essential quality criterions for a nationwide standardized competency assessment. Nevertheless, the opportunity of assessing students' competencies in mathematics education was demonstrated. As Hoogland and Tout (2018) claimed, this study did not tend to reduce contents or competencies, but instead focused on enhancing assessment and innovative items.

REFERENCES

Chituc CM., Herrmann M., Schiffner D., Rittberger M. (2019) *Towards the Design and Deployment of an Item Bank: An Analysis of the Requirements Elicited*. In: Herzog M., Kubincová Z., Han P., Temperini M. (eds) ICWL 2019. LNCS 11841.

Csapó, B., Molnár, G., Toth, K. (2009). Comparing Paper-and-Pencil and Online Assessment of Reasoning Skills. A Pilot Study for Introducing Electronic Testing in Large-scale Assessment in Hungary. The Transition to Computer-Based Assessment: Official Publications of the EU.

Drijvers, P. (2018). Digital assessment of mathematics: Opportunities, issues and criteria. *Mesure et Évaluation En Éducation*, *41*(1), 41–66. https://doi.org/10.7202/1055896ar

Drijvers, P., Ball, L., Barzel, B., Heid, M. K., Cao, Y., & Maschietto, M. (2016). *Uses of Technology in Lower Secondary Mathematics Education: A Concise Topical Survey*. Springer. https://doi.org/10.1007/978-3-319-33666-4

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2019). *Preparing for Life in a Digital World: IEA International Computer and Information Literacy Study 2018 International Report*. IEA.

Geraniou, E., & Jankvist, U. T. (2019). Towards a definition of "mathematical digital competency." *Educational Studies in Mathematics*, *102*(1), 29–45. https://doi.org/10.1007/s10649-019-09893-8

Goldhammer, F., & Zehner, F. (2017). What to Make Of and How to Interpret Process Data. *Measurement: Interdisciplinary Research and Perspectives*, *15*(3–4), 128–132. https://doi.org/10.1080/15366367.2017.1411651

Hoogland, K., & Tout, D. (2018). Computer-based assessment of mathematics into the twenty-first century: pressures and tensions. *ZDM Mathematics Education*, 50, 675-686.

IQB. (n.d.). *VERA - An overview*. Institute for Educational Quality Improvement. Retrieved March 12, 2020, from https://www.iqb.hu-berlin.de/vera

Kaptelinin, V., Affordances, in Lowgren et al., The Encyclopedia of Human-Computer Interaction, 2nd Ed, access on 2020-03-12 from https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/affordances.

KMK. (2003). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss*.

Kortenkamp, U., & Richter-Gebert, J. (2004). Using Automatic Theorem Proving to Improve the Usability of Geometry Software. *Proceedings of MathUI*, 13.

Kovács, Z., Recio, T., & Vélez, M. P. (2018). Using Automated Reasoning Tools in GeoGebra in the Teaching and Learning of Proving in Geometry. *International Journal for Technology in Mathematics Education, 25*(2), 33–50. https://doi.org/10.1564/tme_v25.2.03

Pelkola, T., Rasila, A., Sangwin, C. (2018). Investigating Bloom's Learning for Mastery in Mathematics with Online Assessment. Informatics in Education, v17 n2, 363-380

Rölke, H. (2012). The ItemBuilder. A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of world conference on E-Learning in corporate, government, healthcar, and higher education 2012* (pp. 344–353). AACE. https://www.learntechlib.org/p/41614/

Rupp, A., & Vock, M. (2007). National educational standards in Germany: Methodological challenges for developing and calibrating standards-based tests. In *Making it comparable. Standards in science education* (pp. 173–198). Waxmann.

Sangwin, C., Cazes, C., Lee, A., & Wong, K. L. (2010). Micro-Level Automatic Assessment Supported by Digital Technologies. In C. Hoyles & J.-B. Lagrange, *Mathematics Education and Technology-Rethinking the Terrain: The 17th ICMI Study*. Springer US. https://doi.org/10.1007/978-1-4419-0146-0

Sangwin, C. J., Jones, I. (2017) Asymmetry in student achievement on multiple-choice and constructed-response items in reversible mathematics processes. *Educational Studies in Mathematics Education*, 94, 205–222.

Stacey, K., & Wiliam, D. (2013). Technology and Assessment in Mathematics. In M. A. Clements, A. J. Bishop, C. Keitel, J. Kilpatrick, & F. K. S. Leung (Eds.), *Third International Handbook of Mathematics Education* (pp. 721–754). Springer Science+Business. https://doi.org/10.1007/978-1-4614-4684-2