Monitoring automatically gained difficulty rankings with mathematics educational theories and experts

Eva-Maria Infanger, Nilay Aral, Edith Lindenbauer, Zsolt Lavicza

Monitoring automatically gained difficulty rankings with mathematics educational theories and experts

Eva-Maria Infanger[1], Nilay Aral[2], Edith Lindenbauer[3] and Zsolt Lavicza[1]

[1]Johannes Kepler University, School of Education, Austria; eva_maria.infanger@jku.at

[2]University for Continuing Education Krems, Department of Continuing Education Research and Educational Technologies, Austria

[3]University College of Education Upper Austria, Linz, Austria

Automatically difficulty-ranked tasks would benefit technology-enhanced learning in mathematics, opening adaptive testing for a broader audience. How to achieve this goal in a resource-saving way and guarantee high-ranking quality? This paper follows a community approach for calibration based on the Elo-Rating-System and seeks an instrument to monitor gained task difficulty rankings automatically. Thus, rankings of 18 Algebra-tasks, elaborated following Bloom's Revised Taxonomy, Webb's DOK Framework, and Smith & Stein's LCD, are compared to 5 expert rankings and contrasted to empirical solution frequencies from 64 students in grades 11 and 12. A mixed methods approach will guide the decision for a monitoring instrument for the automatic calibration process implemented in an open test- and trainings-platform based on the GeoGebra classroom containing final exam topics, providing formative assessment and sustaining bridge courses in the STEM fields.

Keywords: Adaptive testing, difficulty level, task complexity, formative assessment, self-directed learning.

Introduction

Numerous research results in psychology and education direct at personalizing learning experiences with care (Chaudhry & Kazim, 2021; Clark-Wilson et al., 2020; Lameier et al., 2018) and show great interest in technology-enhanced learning (Herfort et al., 2023). Meeting the derived requirements in teaching and learning reality is difficult in many directions, which is due to the low number of teachers and the difficulty to pay equal attention to everybody in the learning group.

Digitization enables a wide variety of answers to this gap, which are investigated, designed, and researched carefully (Chaudhry & Kazim, 2021; Pohjolainen et al., 2018). The project "Math Skills Testing" (MST) at Johannes Kepler University Linz in cooperation with the University for Continuing Education Krems, for example, is working on an extension of computerized adaptive testing (CAT) by looking for easier access to the required but labour- and time-intensive calibration process of item sets (Aral & Oppl, 2022). It aims at automation and dynamization via the Elo- Rating-System (ERS) (Pelánek, 2016) to open adaptive testing for a broader audience that lacks both the expertise and the resources to do a proper initial calibration, such as teachers, students or institutions interested in self-directed learning. So, the purpose on a meta-level is to minimize the preparation requirements for CAT (Frey, 2020) on the providers' side, hence, making it more flexible and customizable and ensuring a supportive output on the users' side.

The product will be an open digital webtool based on GeoGebra Classroom that sustains bridging courses for studies in science, technology, engineering, and mathematics (STEM) fields, helping

students prepare before starting at the university and collecting diagnostic data for tutors who provide material and courses for the introductory phase. The webpage will provide item sets in the main mathematics topics for final exam in Austria (divided in the six head chapters *numbers and measures, algebra, analysis, geometry, functional relations, and statistics and probability*) and allows the choice between test and training modes (i.e., items presented without/with formative feedback). The prototype can be found at https://quizzes.geogebra.net/.

This paper documents the search for a scientific monitoring instrument for such an open adaptive testing and training tool at the level of final exam and bridge courses for studies in the STEM fields.

Research Situation

The ERS uses the correct or wrong user input, calculates solution probabilities following the algorithm described in more detail below, and keeps working with those empirical results. As the tool shall be used by anonymous individual learners unsupervised, allowing self-directed learning, the input in the system won't be controlled. Hence, it is imperative to monitor the empirical results through mathematics educational theories and experts and supervise the system. A question arises in this situation: How can the empirical results be monitored to optimize quality and expenses? An item set consisting of 20 selected algebraic mathematics tasks implemented as an openly accessible GeoGebra book (https://www.geogebra.org/m/m8sqmjkk; in the German language, 20.01.2023) serves as the basis for the planned examinations. The theory part provides further details about the ERS and the calculation of difficulty rankings, as it poses the counterpart to the investigated monitoring instruments, analysing not user inputs but items themselves qualitatively. Qualitative ways of determining difficulty rankings shall be introduced, such as, analysing tasks with known mathematics educational frameworks, and a more heuristic way by asking experts to rank the items without any further input based solely on their experience (Mauksch et al., 2020).

To distinguish between the fields of item response theory (IRT) and mathematics educational theories, different terms are used following the distinction between complexity and difficulty (Liu & Li, 2012). Hence, task complexity refers to qualitative mathematics educational approaches, whereas item difficulty indicates an empirical description.

Theoretical Background

To enhance the comprehension of the research design, an introduction to the implementation of adaptive item difficulty and a brief overview of the automatic calibration process is provided. Subsequently, we will outline the selected theoretical frameworks for task complexity and the potential role of experts in an accompanying observation process.

Automatic calibration process

The automatic calibration process applied in this project includes the ERS, which is a method for calculating the relative skill levels of players in two-player games such as chess and go. Applying this system to task ranking, each student has a skill level, and each task has a difficulty level, both starting at a certain value and changing after each solved task based on the outcome, skill level, and difficulty. If a high-skilled student solves an easy task, the difficulty will decrease slightly, while if they fail, it will increase by a larger amount. Conversely, if a lower-skilled student solves a difficult

task, its difficulty will decrease by a large amount, while if they fail, the difficulty will increase marginally. Finally, if the task difficulty is suitable for the student's skill level, the increase or decrease will be moderate. The size of the ranking change is calculated using the equations below (Brinkhuis & Maris, 2009).

$$E_{AB} = \frac{10^{(r_A - r_B)/400}}{1 + 10^{(r_A - r_B)/400}} \qquad r'_A = r_A + K(S - E_{AB}) \qquad r'_B = r_B - K(S - E_{AB})$$

In these equations E_{AB} is the probability of solving the task successfully, r_A is the skill level, and r_B is the difficulty. The calculated solution frequencies are used as initial values.

Mathematics Educational Frameworks

The webtool shall include the whole mathematical material of the final exams after secondary school as well as specified subject areas of certain offered bridging courses. Thus, requirements for a monitoring theoretical framework analysing task complexity are:

- applicability to all mathematical topics,
- simplicity so that it is adaptable at all levels and institutions,
- and enabling a ranking providing for quality and verification of the automatically gained difficulty rankings.

The combination of those demands steers in the direction of well-known frameworks analysing (mathematical) tasks on a meta-level. Literature survey and discussions with mathematics educational experts suggest the investigation of the three most mentioned frameworks of Bloom's (revised) Taxonomy (Krathwohl, 2002), Smith and Stein's (1998) Levels of Cognitive Demand, and Webb's (2007) Depth of Knowledge framework. We will study and compare these three frameworks as for their applicability and reliability in comparison with the initial item calibrations.

The **Depth of Knowledge Framework (DOK)** was developed to investigate the level of alignment between curriculum standards, objectives, and assessment items in the respective states in the U.S. and has a basis for comparison among states. For this purpose, specially trained reviewers are requested to classify the items according to the DOK. As a result the framework defines 4 Levels: 1 – *Recall* of a fact, information, or procedure; 2 – *Skill/Concept*: Use of information, conceptual knowledge, procedures, two or more steps, etc.; 3 – *Strategic Thinking* requires reasoning, developing a plan or sequence of steps; has some complexity; more than one possible answer; generally takes less than 10 minutes to do; 4 – *Extended Thinking* requires an investigation, time to think and process multiple conditions of the problem or task, and more than 10 minutes to do non-routine manipulations with more detailed descriptions The classification may seem to be subjective, but tests show a strong enough correlation. Thus, it seems sufficient for this research aim.

The **Levels of Cognitive Demand (LCD)** evolved during Smith & Stein's (1998) search for good mathematical tasks and guidance for selections. Their main aim is to engage students in high-level thinking. During the generation process, experts (pre-service, in-service teachers, and teacher educators) were discussing the classification of given tasks to four different levels of cognitive demands, two levels each with lower- and higher-level demands, resulting in the levels of *Memorization, Procedures without connections, Procedures with connections,* and *doing mathematics*. The following discussions could not always be brought to a consensus, but the

agreement rate for definitions and limitations was sufficiently convincing and the alignment satisfactory. It is to be emphasized that this specific form of task ranking depends on the respective students' age, knowledge, and level of development.

Bloom's revised taxonomy (BRT) (Krathwohl, 2002) was developed by experts to pave the way for students switching universities. An instrument to level different curricula evolved. The revised version splits the classification of a task into "Knowledge" and "Cognitive" dimension, assumes a cumulative hierarchy, and allows the idea of overlapping adjacent levels. In summary, the revised taxonomy results in a matrix (see Figure 1) and thus enables a differentiated classification of task complexity, as 4 knowledge levels x 6 cognitive levels enable 24 different ranks.

Knowledge dimension	Cognitive dimension: Remember (recalling, recognizing)	Understand (exemplifying, classifying, inferring, interpreting, summarizing, explaining, comparing)	Apply (executing, implementing)	Analyze (organizing, attributing, differentiating)	Evaluate (checking, critiquing)	Create (producing, generating, planning)
Factual knowledge • Terminology • Specific elements and details		6				
Conceptual knowledge • Classifications and categories • Generalizations and principles • Theories, models, and structures		2, 11, 17	7	4		
Procedural knowledge • Subject-specific algorithms and skills • Subject-specific techniques and methods • Criteria for determining when to use appropriate procedures			1, 14, 16	9, 10	5, 13	
Metacognitive knowledge • Strategic knowledge • Cognitive tasks, including appropriate conditional and contextual knowledge • Self-knowledge				8	12, 15	19

Figure 1: Bloom's Revised Taxonomy Matrix[1]

Expert rankings

The term *expert* can be defined in various ways, such as a skilful or well-informed person in a specific field, the sociological view as an individual ascribed with the necessary expertise, selected by a community, or a behavioural approach analysing the ability of individuals to make good decisions, (self-)assessment and consequently their accuracy in predictions. Cognitive psychology distinguishes between an absolute view of persons with innate intellectual competences and a relative subscription of expertise as a level of skills that can be learned and trained (Mauksch et al., 2020). Applied to this context, experts' intuition may be a combination of a feeling for numbers (won through assessing tests and exams) and more qualitative attributes regarding experience with task difficulties by observing students' challenges, misconceptions, and struggles.

---

[1] See: https://pltfrmrsrcscdn.sagepub.com/sk/images/sage-encyclopedia-of-educational-research-measurement-evaluation/10.4135_9781506326139-fig17.jpg, 18.09.2023

Research questions

For all three above-mentioned frameworks, experts play a central role in both creation and implementation. Hence, the possibility of an observation instrument consisting of well-chosen experts will be examined as well. So, the following research questions arise:

R1: Do rankings of difficulty of mathematical algebraic tasks gained from theoretical models, experts, and empirical data differ from one another, and what could be possible influencing factors?

R2: Which theoretical model maps the empirical data in the best way?

R3: Which triangulation approach serves the monitoring process of the results of the ongoing automatized calibration via the ERS?

Methods and study design

In the previous sections, the theoretical foundation for obtaining difficulty rankings in quantitative and qualitative ways was outlined. We will use an approach like Morten and Mühling (2022) who compared estimates of the difficulty of programming constructs.

The analysis is based on the mentioned item set, chosen arbitrarily from a pool of about 300 algebra items taken from open educational resource (OER) item pools (e.g., OPTES project: Küstermann et al., 2021; www.aufgabenpool.at), item numbering following the order in the GeoGebra book. The choice of items was restricted by two assumptions: The difficulty ranking had to portray various difficulties (from very easy to very difficult) to open as wide a range as possible. Also, tasks must give a full coverage of algebra topics needed for the final exams.

64 Austrian 11^{th} and 12^{th} grade students from different school types with different curricula and teachers, preparing for the final exams, were presented during regular lessons in the presence of their mathematics teachers with a specification and question or prompt prepared in the typical digitally correctable task formats (Input-Box and Multiple-Choice). Thus, following CAT (Frey, 2020), this item set evaluates the latent trait "the user can do Algebra" on a specific level. One goal of this design is to diversify the sample as much as possible to get an initial insight into how the data may match or differ. After finishing data collection, 18 items remained for analysis in the end (accidentally deleted item no. 18; double item no. 20). The instruction was to solve each item by hand. The pocket calculator could be used for support, but no Computer Algebra System (CAS). Following the conventional way of item calibration, solutions could be either right or wrong, and scores in between were not possible. Depending on the task, a correct solution could mean identifying the right option(s) or putting in the correct number or formula. So, different levels of engagement and processes were necessary for successful problem-solving. The items were presented in a fixed sequence in GeoGebra classroom lessons shared with teachers and students (a login with previously issued access codes was necessary) but could be edited in any order. No feedback was provided during the test, digital inputs could be changed until end of test. Afterwards, lessons were paused, recording the inputs for analysis, the correctness being assessed digitally, that is, correct or wrong was shown by even or odd numbers in the applets after input, which serve as the only results for analysis.

Presenting these items, three steps of data collection have been taken: i) data from students taking the test to gain solution frequencies (Frey, 2020), ii) intuitive ranking by difficulty by experts (Mauksch et al., 2020), and iii) the application of the frameworks presented above. For the expert rankings, five Austrian experts (two females and three males, all of them chosen based on their experience in upper secondary teaching and task design as teachers or teacher educators) were asked to sort the items by difficulty assuming that the items would be solved by students without the help of CAS. It was particularly pointed out not just to think about their own knowledge but to take different learners and student types into consideration. Just after completing data collection, we started analysing data to avoid previous knowledge influencing the data.

Analysis

It is important to emphasize that this analysis samples rankings, and not ratings. Thus, the given values do not correspond to metrics but only to ordinal data. This enables a comparison of all three ranking approaches. A selection of the results can be seen in Table 1. The numbers represent the position of each item in the given item set. The shades for DOK and LCD distinguish the four levels each, and the thick table lines delimit the four levels according to the DOK, the resulting level pools visualizing why it will show the best relation.

Table 1: Items ranked by difficulty (Nos. represent the position in the item set)

Diffic. rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Sol. Frequ.	6	5	11	7	9	14	2	17	4	12	10	16	1	19	3	13	15	8
Mean (ExR)	7	6	1	11	2	9	14	5	17	16	3	10	4	19	12	13	8	15
BRT	6	2	11	17	7	4	1	14	16	9	10	5	13	8	3	12	15	19
DOK	2	6	7	11	1	5	9	14	16	17	3	4	10	12	8	13	15	19
LCD	2	6	7	11	1	5	14	17	4	9	10	12	16	3	8	13	15	19

Four experts gave back a full ranking, and one expert decided on four levels and assigned the remaining 18 items accordingly. To be able to include these values in the comparison in a meaningful way, the values are interpreted as ties and are therefore averaged over the matching interval. All of them provided further explanations for their choices without being asked. The inter-rater reliability of the expert rankings is analysed using Krippendorffs' alpha (Hayes & Krippendorff, 2007). For all five experts combined, it gives the value of $\alpha = 0.73$ and allows only for tentative conclusions. Excluding expert 4 (who let us know afterwards that the exclusion of CAS was overlooked) increases it to $\alpha = 0.77$. Kendall's coefficient of concordance, another assessment for agreement among raters, results in $\alpha = 0.7940$ for all five experts, and $\alpha = 0.832$ without the rankings of expert 4, indicating strong inter-rater reliability. The relations of the solution frequencies for each item to the theory- and expert-based approaches were determined by Spearman's correlation coefficients. Significance was set at $p < 0.01$ using a two-tailed test. All coefficients show a significant correlation. DOK and LCD

are almost identical ($r = 0.947$); only the level limits are slightly different. BRT differs a little more (with $r \approx 0.865$), which can be explained by the greater differentiation of this framework. Triangulating the solution frequencies, DOK shows the highest correlation with $r = 0.849$, followed by LCD ($r = 0.840$). The correlation with BRT is relatively low with $r = 0.696$. Observing the expert rankings, the highest correlation ($r = 0.769$) can be observed with the mean of four expert ranks, as expected (Mauksch et al., 2020). Expert 5 was excluded because of the level assignment.

Results and discussion

In summary, the various difficulty rankings correlate relatively to very strongly. The results of the correlation analysis answer RQ 2 and suggest the use of a theoretical model was well grounded. In particular, the DOK seems to be a suitable tool for triangulation, especially if a few experts/reviewers invest time in a consensus process (Webb, 2007). The requirement of simple application is best met by the LCS, as it can be applied by pre-service teachers and in-service teachers as well.

To answer RQ 3, it will be necessary to identify the specific needs of the ongoing calibration and monitoring process. If a more detailed ranking is needed, it appears to be more sufficient to collect rankings from corresponding experts and derive the mean than to consult BRT, whereas this framework can help to diversify the item set considerably in visualizing the processes required for solving certain tasks regarding knowledge and cognitive dimension.

Since the test platform is intended to bridge the gap between school and university mathematics, it is assumed that everyone has required the necessary knowledge to solve the posed problems. Nonetheless, it must be considered that it is a completely open webpage. Therefore, no control can be exercised over the data input as it cannot be controlled who accesses it and how. Hence, other more subtle observations are of interest: The distinction in four levels was not only made by two out of three theories but also independently chosen by one expert. Table 1 allows the observation that the items stay within the level boundaries for most rankings. Also, some reasons for varying item ranks could be detected by the explanations of the experts, such as using CAS or not or mastering symbolic thinking, which can change the difficulty rankings of items considerably, thus, confront part two of RQ 1. The statement of LCD that high-level thinking and problem- solving can be achieved by starting with tasks of high-level demand will be considered for further proceedings of the project as it may influence the choice of the rank the first test item shall have.

References

Aral, N., & Oppl, S. (2022). A community-approach to item calibration for testing math-skills in engineering. In M. E. Auer, H. Hortsch, O. Michler, & T. Köhler (Eds.), *Proceedings of the International Conference on Interactive Collaborative Learning ICL 2021* (pp. 454–466). Springer. https://doi.org/10.1007/978-3-030-93904-5_46

Bastian, M., & Mühling, A. (2022). Comparing estimates of difficulty of programming constructs. In I, Jormanainen., & A., Peterson (Eds.), *Proceedings of the 22nd Koli Calling International Conference on Computing Education Research* (pp. 1–12). Association for Computing Machinery. https://doi.org/10.1145/3564721.3565950

Brinkhuis, M. J. S., & Maris, G. (2009). *Dynamic parameter estimation in student monitoring systems.* Measurement and Research Department Reports (Rep. No. 2009-1). Cito.

Chaudhry, M., & Kazim, E. (2021). Artificial intelligence in education (AIEd): a high-level academic and industry note 2021. *AI Ethics, 2*, 157–165. https://doi.org/10.1007/s43681-021-00074-z

Clark-Wilson, A., Robutti, O., & Thomas, M. (2020). Teaching with digital technology. *ZDM Mathematics Education*, *52*(7), 1223–1242. https://doi.org/10.1007/s11858-020-01196-0

Frey, A. (2020). Computerisiertes adaptives Testen. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* (pp. 501–525). Springer. https://doi.org/10.1007/978-3-662-61532-4_20

Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*(1), 77–89.

Herfort, J. D., Tamborg, A. L., Meier, F., Allsopp, B. B., & Misfeldt, M. (2023). Twenty years of research on technology in mathematics education at CERME: A literature review based on a data science approach. *Educational Studies in Mathematics*, *112*(2), 309–336. https://doi.org/10.1007/s10649-022-10202-z

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice, 41*(4). 212–218. https://doi.org/10.1207/s15430421tip4104_2

Küstermann, R., Kunkle, M., Mersch, A., & Schreiber, A. (2021). *Selbststudium im digitalen Wandel* [Self-study in the digital transformation]. Springer. https://doi.org/10.1007/978-3-658-31279-4

Lameier, E., Reinerman-Jones, L., Matthews, G., Biddle, E., & Boyce, M. (2018). Motivational assessment tool (MAT): Enabling personalized learning to enhance motivation. In R. Nkambou, R. Azevedo, & J. Vassileva (Eds.), *Intelligent Tutoring Systems* (pp. 88–98). Springer. https://doi.org/10.1007/978-3-319-91

Liu, P., & Li, Z. (2012). Task complexity: A review and conceptualization framework. *International Journal of Industrial Ergonomics*, *42*(6), 553–568. https://doi.org/10.1016/j.ergon.2012.09.001

Mauksch, S., von der Gracht, H. A., & Gordon, T. J. (2020). Who is an expert for foresight? A review of identification methods. *Technological Forecasting and Social Change*, *154*, Article 119982. https://doi.org/10.1016/j.techfore.2020.119982

Pelánek, R. (2016). Applications of the Elo rating system in adaptive educational systems. *Computers & Education*, *98*, 169–179. https://doi.org/10.1016/j.compedu.2016.03.017

Pohjolainen, S., Nykänen, O., Venho, J., & Kangas, J. (2018). Analysing and improving students' mathematics skills using ICT-tools. *Eurarsia Journal of Mathematics, Science and Technology Education, 14*(4), 1221–1227. https://doi.org/10.29333/ejmste/81869

Smith, M. S., & Stein, M. K. (1998). Selecting and creating mathematical tasks: From research to practice. *Mathematics Teaching in the Middle School*, *3*(5), 344–350. https://doi.org/10.5951/mtms.3.5.0344

Webb, N. L. (2007). Issues related to judging the alignment of curriculum standards and assessments. *Applied measurement in education, 20*(1), 7–25.