ZDM – Mathematics Education (2023) 55:1243–1254
https://doi.org/10.1007/s11858-023-01507-1

ORIGINAL PAPER

Designing an instrument to measure the development of techno-mathematical literacies in an innovative mathematics course for future engineers in STEM education

Nathalie J. van der Wal[1] · Arthur Bakker[2] · Paul Drijvers[2]

Accepted: 19 June 2023 / Published online: 29 June 2023

Abstract
Techno-mathematical Literacies (TmL), which are defined as a combination of mathematical, workplace and ICT knowledge, and communicative skills, are acknowledged as important learning goals in STEM education. Still, much remains unknown about ways to address them in teaching and to assess their development. To investigate this, we designed and implemented an innovative course in applied mathematics with a focus on Techno-mathematical Literacies for 1st-year engineering students, and we set out to measure the learning effect of the course. Because measuring TmL is an uncharted terrain, we designed tests that could serve as pre- or posttests. To prevent a test learning effect, we aimed to design two different but equally difficult tests A and B. These were assigned randomly to 68 chemistry students, as a pretest, with the other one serving as posttest after the course. A significant development in TmL was found in the B-pre group, but not in the A-pre group. Therefore, as a follow-up analysis we investigated whether the two tests were equally difficult and searched for possible explanations. We found that test B was indeed perceived as more difficult than test A, but also that students who were assigned B (pre) were previously higher achieving than A (pre), and a sound mastery level of basic skills that ground the higher-order TmL seemed necessary. Furthermore, as TmL are very heterogenous by nature, some of them are easier learned and measured than others. Based on the results, we propose ways of testing TmL, which should be validated in future research.

Keywords Techno-mathematical literacies · Mathematics education · Assessment · STEM education · Design research · Validation

1 Techno-mathematical literacies in higher STEM education

The professional practices in which science, technology, engineering, and mathematics (STEM) are used have changed over the last few decades because of tremendous changes in available knowledge and digital technology (Duderstadt, 2010; Kent & Noss, 2001). Nowadays, most calculations are performed by computers, and mathematics behind the interfaces are therefore less visible and transparent (Hoyles et al., 2002; Williams & Wake, 2007). In general, mathematical application skills beyond pure mathematical knowledge have been recognised to be increasingly important (FitzSimons, 2002).

To prepare for these new practices, integrated STEM curricula with a broader range of mathematical skills are needed. In this study, we focus on Techno-mathematical Literacies (TmL), which are defined as a combination of mathematical, workplace and ICT knowledge, and communicative skills. Examples of TmL are the ability to interpret abstract data and having a sense of number and a sense of error (Hoyles et al., 2010).

However, how to stimulate and assess the development of TmL in higher STEM education curricula is yet unknown. We, therefore, started a project to foster students' TmL in higher STEM education and integrating mathematics and science. Starting with an interview study in a range of STEM

✉ Nathalie J. van der Wal
n.j.v.d.wal@tue.nl

[1] Eindhoven University of Technology, Eindhoven, Netherlands

[2] Utrecht University, Utrecht, Netherlands

practices, we identified seven commonly used TmL categories. With these TmL as central learning goals, we designed an innovative course in applied mathematics for 1st-year STEM students with the approach of Design-based Implementation Research (Fishman et al., 2013). This course was intended as a first step in developing TmL as long-term educational goal of advancing mathematical skills in higher education, with the workplace as end-goal. The new course, which included group work on applied cases as a core element, was implemented in the curriculum of all majors of the School of Life Sciences and Environmental Technology of Avans University of Applied Sciences and taught by 11 lecturers to thousands of students since 2016 and is still part of the curriculum to this day.

In this paper, we zoom in on the last and the most challenging part of our project, the assessment of TmL. When new learning goals are formulated, it is crucial that these goals can be assessed. Although STEM education aims for interdisciplinary knowledge and skills (Maass et al., 2019), assessment of these heterogeneous skills is scarcely found in the literature. Testing in STEM education is still predominantly discipline-based and assessment tools should be developed for these new objectives (Gao et al., 2020), considering the need for feasibility in assessment at large institutions. Therefore, we evaluated the learning effect of the designed course, by means of a pre- and posttest to measure students' development in TmL. We did not find any previous reported attempts to measure the development of TmL, so we had to design such a test ourselves. The initial research question that guided this study was: What is the learning effect of a course in applied mathematics on students' development of Techno-mathematical Literacies?

Although the project has been successful in various respects (identifying TmL, course design, professional development, and sustainability of implementation), in this last study of our project, we encountered unexpected results. We think these results are worth sharing with others who conduct educational design research, as O'Neill (2012) states: "Principally, it is argued that design researchers must report the failure of designs much more frequently and in a more informative way, and that a critical audience for informative reports of design failure is indispensable to the progress of educational design research" (p. 119). To further investigate our results, we therefore formulated a follow-up question, which reads, what are possible explanations for the contrasting results?

2 Design-based implementation research project

Design-based Implementation Research (DBIR) can provide guidelines for the implementation of design research. This approach aims to develop theoretical insight and at the same time provide practical solutions to complex educational problems. It has emerged to bridge the gap between theory and practice (Bakker, 2018; Van den Akker et al., 2006). It includes iterative, flexible cycles of designing, monitoring, evaluating, and adjusting an intervention, mostly conducted in a team (McKenney & Reeves, 2012). As the implementation of a design is often the most challenging part, and as an expansion of DBR, DBIR has been developed (Fishman & Penuel, 2018). DBIR provides guidelines to support for usability and sustainability of educational interventions. It transcends barriers between the educational disciplines to provide systemic change (Fishman et al., 2013). In our design project, we followed these guidelines to support the implementation of the innovative course in applied mathematics.

2.1 Context of project

The origin of this design project stems in educational challenges. In higher education, STEM subjects are predominantly separated in disciplinary courses, and to move to integrated STEM education, however, there is an increased need for interdisciplinarity or transdisciplinarity (Vasquez et al., 2013). Mathematics courses in higher education are the most "odd ones out". In the Netherlands, they are still taught in a merely abstract manner with little context and almost no connection to other STEM subjects. This was also the case in many schools of Avans University of Applied Sciences in the Netherlands. In the School of Life Sciences and Environmental Technology, where the first author worked as a lecturer, many students were low performing and not very motivated to engage in mathematics. Moreover, most students did not recognise mathematics in their other science courses, and often asked "why do we have to learn this".

Although curriculum changes are not easy in large universities, with support of management, we were given the opportunity to design and implement a new course applied mathematics for 1st-year STEM-students with four colleagues, with the aim of integrating mathematics and science. To develop TmL at a professional level, this course was intended to function as the first step in a learning trajectory. The start in this project was to identify the TmL used by STEM professionals by means of an interview-study.

Table 1 Seven TmL categories that engineers teported to use in their work (from Van der Wal et al., 2017)

	TmL category	Description
1	Data literacy	The ability to analyse and interpret technical data and graphical representations, draw conclusions, and take action accordingly
2	Technical software skills	The ability to use professional software, e.g., Excel™, as a calculation tool
3	Technical communication skills	The ability to communicate technical information with colleagues, customers, supervisors, and other parties
4	Sense of error	The ability to check and verify data and detect errors
5	Sense of number	The ability to handle and interpret numbers sensibly
6	Technical creativity	The ability to produce creative solutions to puzzles and problems (by using, e.g., cleverness or experience)
7	Technical drawing skills	The ability to understand and produce technical drawings (by using, e.g., spatial insight)

2.2 Identifying TmL in STEM practices

The rapid changes in the world due to globalisation, digitalisation and automatisation have caused knowledge to expand at high speed over the last decades. Computers and technology-driven machines have taken over calculations from handwork. Because input has to be monitored with great scrutiny and output has to be interpreted sensibly, there is an increased need to be able to understand quantitative data (Gravemeijer, 2013; Levy & Murnane, 2007). Although mathematics plays a central role in engineering, engineers often perceive themselves as using only simple mathematics (Kent & Noss, 2002). Handling and interpreting abstract information have always been a task for highly trained employees, but because of technology, an increasing number of people engages in these challenges (Kent et al., 2000). Furthermore, because work tasks are nowadays far more complex than they have been in the past, division of labour is practiced, and computations are often outsourced to computers and to expert mathematicians and statisticians. In non-routine tasks, the use of ICT can add a certain mathematical invisibility behind the screen or the print-out, and this can be perceived as a black box (Hoyles et al., 2013; Van der Wal et al., 2017). Professionals learn technology mostly by use, and mathematical literacy, analogous to language literacy, is necessary (Kent & Noss, 2001).

The mathematical literacy needed in professional contexts differs significantly from what is typically taught in formal mathematics education in STEM (Bakker et al., 2006). While the latter addresses merely conventional skills, facts, and procedures, learning in the 21st century should integrate knowledge in a problem-oriented interactive curriculum (Fadel et al., 2007). Garfunkel (2011) emphasises the importance of mathematics learned in the context of science as well, and states that a different set of mathematical skills are necessary, which he identifies as mathematical modelling and quantitative literacy.

Table 2 The weekly class hours schedule

1st hour	Introduction/questions with lecturer
2nd hour	Collaborative work without lecturer
3rd hour	Collaborative work without lecturer
4th hour	Feedback hour with lecturer

As for mathematical skills that go beyond mathematical knowledge, several definitions have been introduced, and because of the technology-driven nature of engineering, we chose to focus on Techno-mathematical Literacies (TmL) in our study. TmL are complex skills that are context specific and based on data and go far beyond numeracy and calculations (Bakker et al., 2006). Even for professional scientists, for example, graphs that originate in other technical domains are often misinterpreted (Roth, 2003).

In the first study of the design project, we conducted an interview study to find out which TmL professionals use in STEM. In the Netherlands, graduates of both engineering and life sciences majors in universities of applied sciences are called 'engineers'. As shown in Table 1, our interview study with 14 engineers from a range of STEM domains led to the identification of seven TmL categories, that these engineers often use in combination (Van der Wal et al., 2017). We will now elaborate on these seven TmL categories with examples from the interview study.

2.2.1 Data literacy

Data Literacy is the ability to analyse and interpret textual, numerical, and graphical data to correctly draw conclusions and taking appropriate action. For example, in our interview with a technical writer for manuals for digital chips machine production, their ability to produce insightful graphics was mentioned to be very important.

2.2.2 Technical software skills

This TmL was the most frequently observed category in our study and concerns the ability to use technical software, both general (e.g., Excel™) and domain-specific technical company software. The key skill in this TmL is not (only) to be able to know 'the buttons' of the software, where the computer is experienced as a black box, but often to know what calculations are performed behind the interface, and therefore, the computer being experienced as a white or grey box. Because we found Excel™ to be the most ubiquitous used software, we decided to implement this tool in our new course.

2.2.3 Technical communication skills

This TmL includes the ability to communicate with various parties, e.g., with colleagues and other departments, but also with management, customers, and employees. Many engineers mention this skill to be very important. A license advisor in environmental engineering uses simple and plain language to support mutual understanding. A sales engineer who designs climate ceilings stresses the importance of asking carefully chosen additional questions to customers. Division of labour and collaboration necessitate interacting and communicating more than ever.

2.2.4 Sense of error

The TmL category *sense of error* concerns the ability to detect errors in all kinds of data, which is a very crucial skill, as small errors can have large effects. The license advisor, for example, reads their received reports with scrutiny, and has to detect conspicuous details.

2.2.5 Sense of number

The ability to handle numbers (but also symbols and formulas) is essential for engineers in all STEM domains. This TmL has an obvious overlap with the previous TmL *sense of error*, but also often combines with *technical software skills*. A technical writer with a background in electrical engineering mentioned the importance of handling units and knowing the difference between milli and micro.

2.2.6 Technical creativity

The sixth TmL encompasses a combination of cleverness, experience and puzzle-solving abilities, especially for professionals who design. A mechanical engineer of large cooling systems mentions a lot of puzzling, boggling, and calculating in their work. Although it is not mathematically difficult, there are a lot of variables involved, and it is just a lot of puzzling with formulas.

2.2.7 Technical drawing skills

The seventh and last TmL that we encountered in our interview study is *technical drawing skills*, which entails the ability to understand and interpret technical drawings, and for some, also producing them as well. For this ability, spatial insight is a critical component, as a permit advisor in the domain of environmental engineering states to need to interpret technical drawings, to understand how things look from the top, side, and front.

2.2.8 Combination of TmL categories

As mentioned before, TmL categories are often combined, as tasks of engineers are often complex. A sales engineer who sells stabilisation fins for yachts, combines *sense of number* with *sense of error*, when checking numerical data of yachts. But using his Excel™ calculation tool with endless formulas, he integrates TmL categories *data literacy*, *technical software skills*, and *technical communication skills*, when thinking on, and discussing with colleagues about the offers of the competition. The technical writer of chip machine manuals needs both *data literacy* and *sense of error* when checking received input. The permit advisor also needs *sense of number* (how much is this variable) and *data literacy* (how does that variable work) in interpreting the technical drawings.

2.3 Course design

With the identified TmL as central learning goals, a new course for 1st year STEM-students named *Applied Mathematics* was designed. The course was collaboratively developed by an interdisciplinary design team of four lecturers consisting of a mathematician (the first author), a chemical engineer, an electrical engineer, and a built environment engineer. These lecturers provided technical contexts and their professional experiences. Cases were created on several topics in the domains of life sciences, electrical engineering, and built environment (e.g., electrical feed of Hall-sensor, linearisation of temperature measurements of a thermistor, temperature balance in a house). In every section of the cases, specific TmL are addressed. The premises of the course were determined based on the input of literature and the interviews with the engineers from the previous study (Van der Wal et al., 2017). As the engineers stressed the importance of application, and TmL need context by nature, we decided to build a problem-based and technology-enriched course of applied mathematics.

2.3.1 Inquiry-based learning

With the aim of using TmL as a central learning goal in an applied mathematics course, we formulated new learning goals which also asked for new pedagogy. We decided to use the approach of inquiry-based learning (IBL). Inquiry is playing an increasing role in science education, as it mimics the patterns of science practices. It stimulates students to acquire and apply science concepts (Linn et al., 1996). IBL is defined as a student-centred approach to stimulate critical thinking, problem-solving and developing an investigating mindset (Anderson, 2002; Chu et al., 2017). Teaching according to this approach involves process-focused questions, while students are engaged and learn actively.

2.3.2 Course structure

With the goal of developing TmL over four years of higher education, our aim in the first-year course was to start this process with using applications for developing TmL, but also laying a strong base of mathematical knowledge and skills. The pragmatic approach we settled on as a design team was to design two parallel but aligned learning tracks: an abstract track with basic mathematics and an applied track with context-rich cases. For the abstract, theoretical track, the software of ALEKS™[1] was used, on which students could work individually, outside class, and at their own level and pace. For the applied track with TmL as goal, we chose to use Excel™ for the technological part, because in our interview-study, we found this tool to be the most used software, and students need Excel™ in many other courses. Students worked collaboratively in groups of 2 or 3 on guided cases during classes. The topics were aligned; for example, linear and quadratic functions were addressed in the first two weeks of the course in both tracks. Because the TmL category of *technical drawing skills* is not used in the technical domains of chemistry and biology, this TmL category was not a learning goal in these cases.

The class schedule consisted of 4 h of classes each week, see Table 2. The first hour was dedicated to introduction of the cases and some instruction. Subsequently, the students worked on the cases for two hours, without the lecturer being present. The fourth hour was dedicated to feedback. Students presented their work to the whole class, or the lecturer visited each group separately. The lecturer adopted the philosophy of IBL to discuss the solutions of the students to the case-problems. In focusing on the approach of problems, rather than solutions, the lecturer tried to foster inquiry-based thinking and help students obtaining knowledge for themselves. Classroom or group discussions can stimulate understanding and competence for complex skills (Nathan & Kim, 2009), such as TmL. Formative assessment was performed in the weekly feedback hours, and the course was concluded with a summative computer-test with TmL items in the context of the cases. Finishing the ALEKS track was also a summative requirement to pass the course.

Because the new course was to be implemented in the curriculum of the School of Life Sciences and Environmental Technology, we eventually chose to use cases specifically for the realms of chemistry and biology, with specific elements of other courses in the curriculum to stimulate recognition of mathematics in sciences and vice versa. The cases were developed to foster the use of six TmL categories. The course contained seven weekly classes, in which every case was worked on for a fortnight. A total of three cases was designed, with the seventh, and last class, dedicated to questions. The first case addressed the topic of pH in solutions and weak acids. In the second case, students worked on bacterial growth and the third case consisted of derivatives and antiderivatives, mainly focusing on qualitative understanding. An extended description of the course can be found in Van der Wal et al. (2019).

2.4 Professional development and implementation

After the design of the course itself, we moved on to scale up and implement the new course in all majors of the School of Life Sciences and Environmental Technology. Because of the new learning goals and pedagogy, this required professional development for the lecturers involved. With the help of one member of the design team and an external process coach, a PD trajectory was developed and conducted, and trained 11 lecturers in the first year to teach the new course. Thousands of students have followed the course in the years following until the present day. The reader can find an extended description of all successes and challenges of this implementation process and PD track for lecturers in Van der Wal et al. 2021.

As a last step of the project, we evaluated student learning in the new course by means of a pre- and posttest. This part is the subject of the rest of this paper.

3 Method

This study involves two phases, a phase 1 in which two tests A and B were developed, validated, and administered with the course as treatment to answer the first research question. In phase 2, we zoom in on the mixed results, to find explanations for them and address the follow-up research question through evaluating the tests. To elaborate on the two phases, we carried out a series of research activities (RA). An overview of these activities is shown in Table 3. Activities 1–11 refer to the first phase, addressing the initial research question of the learning effects. Activities 12–16 concern the follow-up research question in the study's second phase. These research activities are explained in more detail in the next subsections.

[1] Assessment and LEarning in Knowledge Spaces is a Web-based, artificially intelligent assessment and learning system. ALEKS uses adaptive questioning to quickly and accurately determine exactly what a student knows and doesn't know in a course. ALEKS then instructs the student on the topics (s)he is most ready to learn. As a student works through a course, ALEKS periodically reassesses the student to ensure that topics learned are also retained.

Table 3 Overview of research activities

	Research activity	Description	Gain insight into
1	Test design	Designing test items in co-design with students	feasibility
2	Validation	Discussing test items with TmL experts	concept validity
3	Redesign	Adjusting test items and assigning to two tests A and B	content validity
4	Validation	"Thinking-aloud" session with 4th -year chemistry student	construct validity and feasibility
5	Redesign	Adjusting test items regarding language, errors, and number of items	content validity
6	Administering pre- and posttests	Conducting pretest with 68 and posttest with 62 students in 30 min	validity and feasibility
7	Grading	Grading tests by the researcher	test scores
8	Data analysis	Checking for internal consistency of the items with measures of classical test theory.	criterion validity
9		Performing *t* tests to compare pre- and posttest results	content validity
10		Performing one-way ANCOVA to check influence of lecturers.	construct validity
11		Compare P values to investigate development for different TmL categories	content validity
12	Follow up analysis	Performing *t* test to compare marks from course chemical calculation with assignment test A or B as pretest	construct validity
13		Performing *t* tests to compare test A and B	construct validity
14		Performing *t* tests to compare scores on posttest with marks for summative test.	construct validity
15	Redesign proposal	Mixing items within tests to spread missing values	
16		Mixing matching items between test A and B and detect discriminating items to standardise difficulty	

3.1 Design and validation of TmL tests

To measure the learning effect of the new course in applied mathematics, we developed in research activity one (RA1) two tests that both could serve as pre- or posttest of Techno-mathematical Literacies. As the course was part of the regular curriculum, a control group was not possible; moreover because of the new learning goals comparison with results from previous years would not make sense. Because testing of TmL has not been done before, there was no material available to build on, so we had to design such a test ourselves. A co-design team consisting of the lecturer-researcher (first author) and students of Electrical Engineering at Fontys University of Applied Sciences was formed to align more closely to the lifeworld of students and assure feasibility. They were asked to get acquainted with relevant literature and create test items for extracurricular study credits. The lecturer-researcher also trained the students with the concept of TmL with extensive use of examples, and how to apply TmL in test items. They developed these items using as many TmL categories as possible, and to implement these in non-related, non-chemical contexts to avoid treatment-inherency, because with TmL, we aim for transfer beyond the exact tasks used in the course (Cheung & Slavin, 2016). For example, one test item involved an ant that can carry a multiple of its own body weight. The TmL category *sense of number*, which includes numbers, symbols, and formulas, is the TmL which most resembles the "standard" way of mathematics test items, and although we intended not to measure this TmL too often, we see that it is required in many items, as can be seen in Table 4.

Subsequently, in RA2, the pool of potential test items was discussed with two TmL experts, with the aim to validate whether the items indeed measured TmL (concept validity). After further adjustments based on their feedback (RA3), a voice-recorded "thinking aloud" session with a 4th-year Chemistry student of Avans University of Applied Sciences was conducted (RA4). Some possible language confusions were detected, and it appeared that for a half-hour test, we needed to decrease the number of test items. In RA5, the test items were divided in two sets, A (Appendix 5) and B (Appendix 6), by the lecturer-researcher with approximately the same distribution of topics and TmL categories. Subsequently, a grading scheme for both tests was designed by the lecturer-researcher. Points per test item were based on the number of steps a student had to take to come to a solution. Test A yielded a total of 33 points and test B 28 points, both normalised to percentages (0−100). We estimated that 13 items for test A versus 12 items for test B would take approximately the same amount of time for students.

3.2 Participants and procedure

The tests were administered to all 1st-year Chemistry students (N = 68, 38 female and 30 male, aged 17–24) between April and June of 2019, during the third cycle of the implementation (RA6). The 68 students were randomly divided over two groups with two lecturers, and therefore, for the pretest, one group was assigned to test A, and the other group to test B. To prevent the previous mentioned test learning effect, we assigned the 62 students that were present during the posttest to the other test.

The pretest was conducted at the start of the first class of the course in applied mathematics. All students signed a consent form and were given exactly 30 min to perform the test. Because the seventh and last class was not obligatory and functioned as a question session, we decided to conduct the posttest during the sixth week. As participating in this test relied on the goodwill of the students, we had to keep the tests short and schedule it at the beginning of the sixth class rather than at the end, so the students would not be tired already. The mathematical content of the sixth class, therefore, was not covered yet, and this could have influenced scores on the posttest, which we discuss further in the results section.

Although the course was computer-based, as students used Excel to work on the cases, the tests were conducted with pen, paper, and calculator, due to practical constraints. The test had to be administered during class-time, in a normal classroom without computers. This decision was based on the premise that, although Tml connect mathematics and technology, they are not skills of handling 'the buttons' of complex software. They have a base in mathematics and are broader than technology or ICT. Moreover, problem solving computationally is not just a digital skill, but rather a mental skill. TmL can, therefore, to some degree also be learned and tested "unplugged", as is practised in the development of computational thinking (Caeli et al., 2020; Kallia et al., 2021; Rodriguez, 2017). We included "Excel" test-items to test TmL *technical software skills.*

The students were asked to fill in how much time of the half-hour was left when they had finished, to gain insight into how much time students need for doing the two tests (feasibility). The mark that students received for the course in applied mathematics was unrelated to the pre- and posttest scores. To pass the course, they had to master 90% of the assigned topics in the learning track of ALEKS™ as a prerequisite to receive a mark for their summative digital test, two weeks after the posttest. With TmL questions in the contexts of the cases, the students worked on during the course, summative test items were more familiar to them than the pre- and posttest. As a measure of item difficulty, we use the P value, the average correct score of all students (not to be confused with *p* value to estimate statistical significance).

3.3 Scoring

After conducting both pre- and posttests, the tests were divided and scored by the lecturer-researcher and the other teaching-lecturer. Because the test was paper based, all information about the students and which test they were assigned to was visible for the lecturers. Based on the grading scheme, both lecturers graded the tests with frequent peer consultation. After intensive discussion of the grading procedure with the other lecturer, the lecturer-researcher did all the grading to ensure consistent application of the grading procedure (RA7). To be able to judge whether items were too difficult or whether there were too many items (feasibility criterion), we distinguished the assignment of a code of zero points or a code 999. The code 999 was used for missing data; when nothing was filled in, when a question mark or an "I don't know" was filled in, or when the question was rewritten but no answer was given. A score of zero points was used when something was tried, but the answer was wrong. Zero points were also assigned when something was tried wrongly but crossed out; this was not considered as missing data.

3.4 Data collection

The scores on the items of the two tests were the main data collected for this study in phase (1) However, after the unexpected contrasting results of the tests to measure the development of TmL from pre- to posttest, and further research was necessary, extra data was collected for phase (2) First, we wanted to test whether the random assignment of test A and test B among the students led to groups with about the same average level in mathematics. Therefore, for RA12, we collected the marks of the students of a course on chemical calculation, that the students followed earlier that academic year and compared these marks, normalised to percentages (0−100) with the students' assignment to test A or B as pretest. Secondly, for RA14, we collected the marks of the summative digital test, also normalised to percentages (0−100) and compared these marks with the scores on the tests A and B as posttest, to find out whether test A and B have the same level of difficulty.

4 Results

In Sect. 4.1− 4.5, we present the results regarding phase 1, answering the initial research question on the learning effect concerning TmL from pre- to posttest, as described in research activities 8−11. Section 4.6 − 4.8 contain the results on the follow-up research question on explanations for the test results, corresponding to research activities 12−14 (phase 2).

4.1 Internal consistency

Validity is not a singular concept, but in essence boils down to measuring what one intends to measure (Borsboom et al., 2004; Hoogland et al., 2016). The Standards of Educational and Psychological Testing define validity as "the degree to which evidence and theory support the interpretation of test scores for proposed uses of tests" (AERA et al., 2014, p. 11).

Gardner (1995) emphasises the fact that in a rating scale, when scores are summed up, it is important that all items reflect the same construct and according to Taber (2018), a high value of alpha is not always a good thing, for it cannot ensure that an instrument or scale is unidimensional and a high alpha may even indicate a use of an inefficient level of redundant items. Therefore, interpretation of this value is not unambiguous.

In our context, we focused on the feasibility for students to do the test in time as well as concept and construct validity with the proposed use as pre- and posttest to be able to measure development in the TmL developed by engineering students. Although we did not consider TmL to be a homogenous construct and although we were aware of the discussion among experts about internal consistency, we studied the internal consistencies of the tests to learn about the heterogeneity of the skills we purported to measure.

Although, we conjectured low alphas because of the heterogenous nature of TmL, we see in Appendix 1, relative normal alphas for test A, but lower values for test B.

4.2 P values

In RA8, the P values, which are the normalised average scores per test-item, were calculated, as for test A used as pretest, Appendix 2 shows an average P value of 0.63, which means students did well on this test. In test A (post), which was assigned to the other half of the student population, we see an average P value of 0.65. As for test B (post), with an average P value of 0.59, we have fewer high outliers than in test A (post). This suggests that test B was more difficult than test A. Concluding, we see a small increasing P value, and therefore a small learning effect from pre to post in both tests.

In RA8 we also calculated the percentages of missing values. Although we shortened the tests after the aloud session with the test student, the percentage of items that were reviewed as missing value increased towards the end of the tests – see Appendix 3 – which suggests that the tests were too long to finish in time for many students. Test B shows more missing values in general, which could indicate that this test was more difficult for the students. Furthermore, item 4b in test A and item 1b and 2 in test B, show high values, which are probably more difficult items.

4.3 Pre- and posttest compared

In comparing the results on the pre- and posttest in RA9, we performed three paired *t* tests. The *t* test on the total scores for pretest ($M = 58.38$, $SD = 18.17$) and posttest ($M = 61.41$, $SD = 19.11$) did show improvement, but not a significant increase, $t(60)$ = -1.059, $p = .294$, 95%CI[-8.760, 2.694], $d = 0.16$. Scores on test A (pre) ($M = 62.69$, $SD = 17.61$) and test B (post) ($M = 58.63$, $SD = 20.14$) showed a decrease in results, but not significant, $t(31) = 1.003$, $p = .324$, 95%CI[-4.202, 12.327]. However, scores on test B (pre) ($M = 53.62$, $SD = 17.87$), and test A (post) ($M = 64.48$, $SD = 17.74$) did improve significantly, $t(28)$ = -3.047, $p = .005$, 95%CI[-18.165, -3.559], $d = 0.61$. Wilcoxon signed rank tests showed similar results. Concluding, we see an overall small, not significant, development in learning TmL, because the results of B (pre) – A (post) are contrasting with A (pre) – B (post). These results indicate that test B was most likely more difficult than test A.

As mentioned before, the posttest was scheduled in the sixth week before all the mathematical content was provided to the students. The topic of anti-derivatives was not covered yet, and this topic was addressed in item 3 of both tests. We therefore excluded this item and performed the three paired *t* tests again, but this did not show a different result. Because of this result, we decided to include this item in further analysis.

4.4 Lecturers compared

Because the course was taught by two lecturers to three classes and to check whether a difference in score pre- to posttest could be due to lecturers' influence, a one-way ANCOVA was conducted in RA10 to determine a difference between lecturer 0 ($M = 65.12$, $SD = 20.59$) and lecturer 1 ($M = 60.32$, $SD = 23.09$) on the scores of the students on the posttest controlling for the pretest. This difference was not statistically significant ($F(1,58) = 0.604$, $p = .440$).

4.5 Different TmL compared

As to see how P values differ when we match them with their corresponding TmL categories 1–6, we grouped the items in RA11 for both tests A and B as shown in Tables 4 and 5. This distribution shows that test items often include multiple TmL categories. The largest progress, however, is found in the two TmL *technical software skills* and *sense of error* in test B, which can be explained by the extensive and explicit presence of these TmL categories in the course. The TmL category *sense of error* is the only TmL that increased in both tests. TmL category *technical communication skills* was only addressed in item 3b in test A and in item 3a in test B, part of items that deal with theory students had not received yet. We see that some TmL categories, such as *sense of error* and *software skills* are quite concrete, demarcated, and probably more easily learned and measured, whilst others, for example *data literacy* and *technical creativity* are broader concepts and more difficult to be learned and measured.

Table 4 Distribution of test items over TmL categories in test A and B

Test A	TmL	Item												
	1	1a			2a		3a	3b	4a	4b	5		7a	7b
	2								4a	4b				
	3							3b						
	4					2b			4a	4b				
	5		1b	1c	2a	2b	3a		4a	4b	5	6	7a	7b
	6	1a					3a	3b		4b	5			
Test B	TmL	Item												
	1	1a	1b		3a	3b	3c			5		7a		
	2							4a	4b					
	3				3a									
	4							4a	4b		6			
	5	1a	1b	2		3b		4a	4b	5	6	7a	7b	
	6			2			3c			5		7a		

4.6 Marks on course in chemical calculation compared with test A or B as pretest

In comparing the marks of the students on the course in chemical calculation with being assigned to test A or B as pretest (RA12), we found, in an unpaired *t* test, a difference between test A ($M = 71.94$, $SD = 14.83$) and test B ($M = 78.35$, $SD = 13.82$), although marginally not significant, $t(63)$ = -1.779, $p = .077$, 95%CI[-6.414, 3.565]. It seems that students who performed test B as a pretest achieved higher on the course chemical calculation than the students who were assigned to test A as pretest. This could be an explanation as to why the B (pre) – A (post) students showed TmL development, and group A (pre) – B (post) did not. The decrease of 0.21 in test B, versus an increase of 0.09 in test A, as seen in Table 5, could be attributed to the fact that test B (post) was assigned to the lower achieving students. This might also explain why we see a development in four TmL categories in test A and in only two in test B. Because test B (pre) students were higher achievers, and the difference between the scores in test A (post) and in test B (pre) is significant, we can also assume that these students have learned more in this course. For higher-order skills, such as TmL, the learning effect often depends on the academic level.

4.7 Tests A and B compared

In RA13 we compared test A and B with independent *t* tests to investigate whether they indeed differ in difficulty, as was suspected by the results on the development from pre- to posttest, as shown in Sect. 4.3. In comparing test A (pre) ($M = 62.46$, $SD = 17.46$) with test B (pre) ($M = 53.97$, $SD = 18.05$), we see that students on test A scored higher, although marginally statistical not significant, $t(66) = 1.971$, $p = .053$, 95%CI[8.487, 4.306]. We found the same in the posttest; although scores on test A (post) ($M = 64.48$, $SD = 17.74$) are higher than on test B (post) ($M = 58.63$, $SD = 20.14$), it is not a significant difference, $t(59) = 1.200$, $p = .235$, 95%CI[-3.910, 15.626]. Three Mann-Whitney tests confirm these results. Boxplots of the scores on both pretests and posttests are presented in Table 4. Altogether, there is reason to suspect that test B is more difficult than test A.

4.8 Posttests scores and summative digital test marks compared

To investigate whether test B was indeed more difficult than test A, the marks of the students' summative test are compared with the scores on the posttest in RA14. A paired *t* test, comparing the test A (post) results ($M = 64.48$,

Table 5 P values and differences of TmL categories in test A and B

		A			B		
TmL		pre	post	difference	pre	post	difference
1	Data literacy	.72	.71	− .02	.55	.52	− .03
2	Technical Software skills	.57	.56	− .01	.33	.57	.24
3	Technical communication skills	.61	.70	.09	.80	.58	− .21
4	Sense of error	.54	.61	.06	.38	.62	.24
5	Sense of number	.55	.62	.07	.57	.54	− .03
6	Technical creativity	.58	.66	.09	.59	.55	− .04

$SD = 18.39$) with the summative test marks ($M = 71.48$, $SD = 16.81$), showed not to be significant, $t(26)$ = -1.723, $p = .097$, 95%CI[-15.352, 1.352]. The paired t test, comparing test B (post) results ($M = 58.63$, $SD = 20.14$) with the summative test marks ($M = 75.56$, $SD = 12.78$), did show to be significant, $t(31)$ = -4.200, $p < .001$, 95%CI[-25.163, -8.172], $d = 1.00$, which indicates that test B is indeed more difficult than test A.

As content experts, it is not clear to us why test B is perceived as more difficult. Especially item 7b of both tests, in which a growing factor is asked, seems equally difficult. However, item 7b of test A (Appendix 5) showed an increase in P value of 0.12, but test B (Appendix 6) a decrease of 0.14. Therefore, we presented both items to a non-participating colleague in the School of Life Sciences and Environmental Technology to judge these two items. He stated that he could not see any difference in difficulty in both items. Moreover, he conjectured that the item in test A could be perceived as more difficult because of the use of more formal mathematical language. An explanation for this could be, again, the fact that the higher achieving students were assigned to test A as posttest. It is also often the case that last items of a difficult test score rather poorly because students get tired, lack time, or give up.

5 Conclusion and discussion

In this paper, we first described our design project, which included an interview study to identify used TmL in a variety of science and engineering domains. With these TmL as learning goals, we designed and implemented an innovative course in applied mathematics in which abstract mathematics is combined with science contexts of other courses in the curriculum. As a last step, we tried to evaluate the course with the research question of what the learning effect is of a course in applied mathematics on students' development of Techno-mathematical Literacies.

Unfortunately, we can only provide a mixed answer to this question. On the one hand, we do see progress from pre- to posttest scores, which suggests a positive learning effect on TmL skills. On the other hand, the answer is preliminary, as the progress is not large, and the picture is somewhat distorted by differences between the two test that were used and the slight initial differences between the two test condition groups.

The tests A and B were randomly assigned to the 68 participating students. For the posttest, the two tests were reversely assigned to the 62 present students as to prevent a test learning effect. In analysis we compared pre- and posttest scores and found a mixed result; the scores of the students from test B (pre) to test A (post) increased significantly, but the scores of students from test A (pre) to test B (post) did not. Furthermore, using an one-way ANCOVA, we found no evidence for a difference in lecturers.

We did, however, report a progression on scores in certain TmL categories. TmL *sense of error* showed a large increase in scores between pre and post in both pairs of tests. For TmL *software skills* we see a large improvement from test A to test B. These TmL categories are rather concrete and demarcated, and we think they are easier to learn and measure than a TmL category such as *technical creativity* or *data literacy*, which are more complex and probably need more time to develop.

From these results, we suspected that test B was more difficult than test A. Therefore, we investigated the follow-up question of what possible explanations there might be for the contrasting results. To answer this question, we first investigated if the two groups of students were indeed equivalent. To this end, the scores of a previous course in chemical calculation that students followed earlier that year were compared with the assignment to test A or B as pretest. We found that students who were assigned to test B as pretest were higher achieving than the students with test A as pretest, although marginally statistically insignificant. This could be an explanation as to why the higher achieving students showed more development from pre- to posttest, but also indicated that test B could be more difficult than test A.

Subsequently, we compared test A (pre) to test B (pre) and test A (post) to test B (post). Although an unpaired t test did not show a significant difference between test A and B, and the lecturers could not recognise a difference in difficulty, we decided to compare posttest scores with the summative test marks of the course, and with these paired t tests, it was proved that students scored significantly less on test B than on test A. We cannot provide an explanation for this fact, it was unexpected and as content experts, we do not recognise this difference. The test items distribution over test A and test B, therefore, should be re-evaluated, and discriminating items should be identified, so as a next step, the tests A and B should be equivalated with respect to difficulty.

As a next step, we plan to randomise test items both within (RA15) and between the two tests A and B (RA16). This will ensure that the highest percentages of missing values are not concentrated at the end of each test, and more importantly, as mentioned above, by mixing matching items from both tests and detecting discriminating items, we can equivalate the tests in difficulty.

Before we discuss these conclusions, we first have to point out some of the study's limitations. On the practical side, we had to deal with a limited sample size and number of test items, because we could not spend more than half an hour to perform the test, as class-time is limited, and students participated voluntarily in this study.

In conclusion, it is clear that we need an upscaling to complex, situational, and contextual learning goals for the aim of integrated STEM education. Designing and implementing of new curriculum elements can have challenges, however, measuring of students' development of certain learning goals is even more difficult. We noticed, that, in measuring TmL, we entered new territory, as this has not been done before and no material was available. In general, designing test items for complex skills is not an easy task (Drijvers et al., 2019). The measurability of TmL, however, was completely unclear and this study took a first step and contribute to the knowledge of testing TmL.

With this, we mention the discussion about "measure what you find important" versus "finding important what you measure" and the challenge not to focus on just teaching what you can measure (Collins, 2017). Can TmL, and other, higher order, integrated, or interdisciplinary skills be measured by psychometric testing of singular topics, or should they, for example, be assessed via qualitative methods? During the implementation of the new course, we tried oral assessments for summative testing. This was promising in the respect of testing complex learning goals, although practically too time-consuming and calibration among colleagues proved to be challenging. After all, in regular education, one has to deal with the grammar of schooling, in our case 1st-year courses with large numbers of students.

Furthermore, one can wonder whether complex learning goals can be developed in just one course or might need more time to develop. Our project has shed some light on poignant issues that may be more general. As TmL can be interpreted as end-goal in a mathematics learning trajectory from primary education to disciplinary secondary education, via higher education, to STEM workplace, we need to focus on a curriculum as a whole, in which an introductory course of applied mathematics is, of course, just one step. To aim for broad scale interdisciplinarity or transdisciplinarity, a larger curriculum change is necessary than this local opportunity of five people and one course. The role of mathematics in advancing integrated STEM education is, therefore, a topic for further research. However, as the new course is still a thriving part of the curriculum, the goal of DBIR as sustainable beyond the end of a project is met. Many students engage and perform well in the course, and we have not heard the question "why do we have to learn this" anymore.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11858-023-01507-1.

Acknowledgements This work is part of the research program Doctoral Grant for Teachers, with project number 023.009.061, which is financed by the Netherlands Organisation for Scientific Research (NWO). We would like to thank Amina Mokhtar, Jan Roelofs, Joost Giesbers, Ehsan Lakzaei, and Vlad Svarnovics, students from Electrical Engineering at Fontys University of Applied Sciences, for their creative work in designing test items. We would also especially like to thank Edwin Melis, colleague at Avans University of Applied Sciences, for his support and help in conducting the tests.

References

AERA, APA, &, & NCME (2014). Standards for educational and psychological testing, 11–31.

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education*, *13*(1), 1–12.

Bakker, A. (2018). *Design research in education: A practical guide for early career researchers*. Routledge.

Bakker, A., Hoyles, C., Kent, P., & Noss, R. (2006). Improving work processes by making the invisible visible. *Journal of Education and Work*, *19*(4), 343–361.

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071.

Caeli, E. N., & Yadav, A. (2020). Unplugged approaches to computational thinking: A historical perspective. *TechTrends*, *64*(1), 29–36.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, *45*(5), 283–292.

Chu, S. K. W., Reynolds, R. B., Tavares, N. J., Notari, M., & Lee, C. W. Y. (2017). *21st century skills development through inquiry-based learning: From theory to practice*. Springer.

Collins, A. (2017). *What's worth teaching?: Rethinking curriculum in the age of technology*. Teachers College Press.

Drijvers, P., Kodde-Buitenhuis, H., & Doorman, M. (2019). Assessing mathematical thinking as part of curriculum reform in the Netherlands. *Educational Studies in Mathematics*, *102*(3), 435–456.

Duderstadt, J. J. (2010). *Engineering for a changing world*. Springer.

Fadel, C., Honey, M., & Pasnik, S. (2007). Assessment in the age of innovation. *Education Week*, *26*(38), 34–40.

Fishman, B., & Penuel, W. (2018). Design-based implementation research. In J. D. Slotta, R. M. Quintana, & T. Moher (Eds.), *International Handbook of the Learning Sciences* (pp. 393–400). Routledge.

Fishman, B. J., Penuel, W. R., Allen, A. R., Cheng, B. H., & Sabelli, N. (2013). Design-based implementation research: An emerging model for transforming the relationship of research and practice. *National society for the study of education*, *112*(2), 136–156.

FitzSimons, G. E. (2002). *What counts as mathematics? Technologies of power in adult and vocational education*. Kluwer.

Gao, X., Li, P., Shen, J., & Sun, H. (2020). Reviewing assessment of student learning in interdisciplinary STEM education. *International Journal of STEM Education*, *7*(24), 1–14.

Gardner, P. L. (1995). Measuring attitudes to science: Unidimensionality and internal consistency revisited. *Research in Science Education*, *25*(3), 283–289.

Garfunkel, S., & Mumford, D. (2011). *How to fix our math education* (p. A27). The New York Times.

Gravemeijer, K. (2013). Mathematics education and the information society. In A. Damlamian, J. F. Rodrigues, & R. Sträßer (Eds.), *Educational interfaces between mathematics and industry* (pp. 279–286). Springer.

Hoogland, K., Pepin, B., Bakker, A., de Koning, J., & Gravemeijer, K. (2016). Representing contextual mathematical problems in descriptive or depictive form: Design of an instrument and validation of its uses. *Studies in Educational Evaluation*, *50*, 22–32.

Hoyles, C., Wolf, A., Molyneux-Hodgson, S., & Kent, P. (2002). *Mathematical skills in the workplace*. Technology and Mathematics Council: Science.

Hoyles, C., Noss, R., Kent, P., & Bakker, A. (2010). *Improving mathematics at work: The need for Techno- Mathematical Literacies*. Routledge.

Hoyles, C., Noss, R., Kent, P., & Bakker, A. (2013). Mathematics in the workplace: Issues and challenges. In A. Damlamian, J. F. Rodrigues, & R. Sträßer (Eds.), *Educational interfaces between mathematics and industry* (pp. 43–50). Springer.

Kallia, M., van Borkulo, S. P., Drijvers, P., Barendsen, E., & Tolboom, J. (2021). Characterising computational thinking in mathematics education: A literature-informed Delphi study. *Research in mathematics education*, *23*(2), 159–187.

Kent, P., & Noss, R. (2000). The visibility of models: Using technology as a bridge between mathematics and engineering. *International Journal of Mathematical Education in Science and Technology*, *31*(1), 61–69.

Kent, P., & Noss, R. (2001). Finding a role for technology in service mathematics for engineers and scientists. *The teaching and learning of mathematics at university level* (pp. 395–404). Springer.

Levy, F., & Murnane, R. (2007). How computerized work and globalization shape human skill demands. In M. M. Suárez-Orozco (Ed.), *Learning in the global era: International perspectives on globalization and education* (pp. 158–174). University of California Press.

Linn, M. C., Songer, N. B., & Eylon, B. S. (1996). Shifts and convergences in science learning and instruction. In B. Berliner, & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 438–490). Routledge.

Maass, K., Geiger, V., Romero Ariza, M., & Goos, M. (2019). The role of mathematics in interdisciplinary STEM education. *Zdm*, *51*(6), 869–884.

McKenney, S., & Reeves, T. (2012). *Conducting educational design research*. Routledge.

Nathan, M. J., & Kim, S. (2009). Regulation of teacher elicitations in the mathematics classroom. *Cognition and Instruction*, *27*(2), 91–120.

O'Neill, D. K. (2012). Designs that fly: What the history of aeronautics tells us about the future of design-based research in education. *International Journal of Research & Method in Education*, *35*(2), 119–140.

Rodriguez, B., Kennicutt, S., Rader, C., & Camp, T. (2017, March). Assessing computational thinking in CS unplugged activities. In *Proceedings of the 2017 ACM SIGCSE technical symposium on computer science education* (pp. 501–506).

Roth, W. M. (2003). Competent workplace mathematics: How signs become transparent in use. *International Journal of Computers for Mathematical Learning*, *8*(2), 161–189.

Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, *48*(6), 1273–1296.

Van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (Eds.). (2006). *Educational design research*. Routledge.

Van der Wal, N. J., Bakker, A., & Drijvers, P. (2017). Which techno-mathematical literacies are essential for future engineers? *International Journal of Science and Mathematics Education*, *15*(Supplement 1), 87–104.

Van der Wal, N. J., Bakker, A., & Drijvers, P. (2019). Teaching strategies to foster techno-mathematical literacies in an innovative mathematics course for future engineers. *Zdm*, *51*, 885–897.

Van der Wal, N. J., Bakker, A., Moes, A., & Drijvers, P. (2021). Fostering techno-mathematical literacies in higher technical professional education: Reflections on challenges and successes of DBIR. In Z. A. Philippakos, J. W. Pellegrino, & E. Howell (Eds.), *Design based Research in Education: Theory and applications* (pp. 296–316). Guilford Press.

Vasquez, J., Sneider, C., & Comer, M. (2013). *STEM lesson essentials, grades 3–8: Integrating science, technology, engineering, and mathematics*. Heinemann.

Williams, J. S., & Wake, G. D. (2007). Black boxes in workplace mathematics. *Educational Studies in Mathematics*, *64*, 317–343.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.