13th IFAC Symposium on Advances in Control Education
July 24-27, 2022. Hamburg Bergedorf, Germany

Assessing engineering exercises: a novel taxonomy

Marica Liotino *, **Anja Garone** **, **Steffi Knorn** ****,
Damiano Varagnolo ***, **Emanuele Garone** †, and
Monica Fedeli *

* *University of Padova, Italy*
** *Vrije Universiteit Brussel, Belgium*
*** *Norwegian University of Science and Technology, Norway*
**** *Technische Universität Berlin, Germany*
† *Université libre de Bruxelles, Belgium*

Abstract: Formally describing and assessing the difficulty of learning and teaching material is important for quality assurance in university teaching, for aligning teaching and learning activities, and for easing communications among stakeholders such as teachers and students. We propose a novel taxonomy to describe and quantify the difficulty levels of exam questions and exercises encountered in engineering-related contexts, together with its development and piloting processes. The taxonomy consists of two dimensions which separately and independently describe the difficulty in understanding / explaining and using / applying a content unit. The piloting phase included 10 purposefully selected experts in the field of control engineering, external to the project, that tested its performance, utility, ease of use, and clarity. The results indicate that the users were able to provide consistent and coherent assessments of the difficulty levels of 15 selected exam questions. The paper further discusses suggestions for improvement voiced by the participants in order to promote an even more consistent and coherent assessment of engineering students' mastery of the subject.

Keywords: assessment tool, Systems and Control, higher education

1. INTRODUCTION

This study is part of a wider Erasmus plus research project titled "Face It: Fostering Awareness on Program Contents in Higher Education using IT tools" realized by Uppsala University, Université Libre de Bruxelles, Otto-von-Guericke University Magdeburg, Norwegian University of Science and Technology, and the University of Padua. Among its aims, the project focuses on improving the common understanding of what is taught in courses and what is expected from students. Thus it includes improving the clarity, efficiency and effectiveness of the forms of information exchange among teachers, students and administrative staff. Indeed, anecdotal evidence suggests that exchanging information about courses contents is not trivial: frustration often arises when discovering that different courses teach the same content, assume prior knowledge that has not been provided yet, or the outcomes as intended by the teachers do not correspond with the learning outcomes perceived by the students.

Parts of these problems can be tackled using constructive alignment, a curriculum design approach that seeks to optimize the conditions for quality learning, as well as building a coherent learning environment where teaching methods and assessment practices are aligned with teaching objectives [McMahon and Thakore, 2006].

In this frame, as a first step we seek to contribute towards standardizing how to describe the difficulty level of what is being taught by means of a taxonomy that should ideally enable to index in a valid and objective way how advanced knowledge/skills an assessment material requires. In addition to the creation of a shared language, this tool could support teachers in selecting appropriate exercises for the assessment of the corresponding intendend learning outcomes in their various shades of difficulty.

Indeed, taxonomies have their roots in the curriculum design and development movements, and have their initial focus on sequential structuring and objective assessment of learning. Bloom et al. [1956] first introduced the concept of taxonomy of educational objectives with the aim of reducing the ambiguity of teaching activities, and of organising in a sequential way the process of assessment. The goal was to identify expected behaviours and the required skills for their achievement. Arguably, this taxonomy is the most well known and consists of three domains: cognitive, affective and psychomotor; each domain is divided into categories [Bloom et al., 1956, 1964, Simpson, 1971]. The cognitive domain is also known as *Bloom's Taxonomy* and it is widespread amongst engineering educators as a framework to describe complexity and higher-order thinking [Stotsky, 2017, Mead and Bennett, 2009]. Some scholars have found the taxonomy useful for the design and assessment of

⋆ The research leading to these results has received funding from pedagogical funds at Norwegian University of Science and Technology, from the International Federation of Automatic Control, and the European Community through the Erasmus+ project 2019-1-NO01-KA203-060257 "Face It".

software engineering courses [Britto and Usman, 2015], and to improve the alignment of assessment and learning outcomes and the quality of software engineering teaching [Khairuddin and Hashim, 2008]. Among the strengths, the extensive analysis of test items, its simplicity, and the distinctness of factors of the cognitive domain have been identified [Fuller et al., 2007, p. 156]. Nevertheless, some critiques have been addressed on Bloom's taxonomy. In particular several authors report that it is not suitable for the computing context [Masapanta-Carrión and Velázquez-Iturbide, 2018, Azuma et al., 2003] and does not adequately address the skills and competences needed in engineering [Heywood, 2005, p. 28]. Other difficulties have been highlighted concerning the differentiation of the cognitive activity involved in each category [Fuller et al., 2007, Staffas et al., 2020]. Hence, considering students' cognitive processes can become challenging [Kallia, 2017]. Additionally, this taxonomy offers different interpretations [Johnson and Fuller, 2006, Heywood, 2005, Staffas et al., 2020] and overlaps among categories have been highlighted [Fuller et al., 2007, p. 156] that make some learning goals to fit in more than one category [Masapanta-Carrión and Velázquez-Iturbide, 2018, Staffas et al., 2020]. There are also disagreements in categorizing knowledge, in particular related to higher levels [Azuma et al., 2003, Fuller et al., 2007]. In fact, the applicability of these categories to every module has been problematic [Johnson and Fuller, 2006] and therefore not suitable for undergraduate courses [Ardis et al., 2015, p. 17]. Furthermore, some authors consider the taxonomy not exhaustive enough as it neglects operational knowledge [Azuma et al., 2003] and accordingly propose adding other categories to the taxonomy [Heywood, 2005].

Another taxonomy often used in the field of engineering and computer sciences is the SOLO (Structure of the Observed Learning Outcome) taxonomy [Biggs and Collis, 2014]. The SOLO taxonomy consisting of five levels of knowledge, is based on Piaget's stages of cognitive development, and aims at capturing adult conceptual development. Its intuitiveness and reliability [Stotsky, 2017, Watson et al., 2014], its usefulness to analyze student's knowledge [Watson et al., 2014, Kallia, 2017], and its holistic nature [Fuller et al., 2007] are its strengths. On the other hand, weaknesses have also been reported in the literature such as imprecision on specific concepts learned [Watson et al., 2014, Staffas et al., 2020] and lack of usage experiences in the field [Fuller et al., 2007].

To the best of our knowledge, and on the basis of our field experience, none of the existing taxonomies are suitable for defining an objectively and reproducible way for indexing exam questions according to their subject, complexity and difficulty. The main perceived issues are the *subjective interpretability* associated to the existing taxonomies and the inability of most taxonomy to describe content having both a "theory" and a "practice" side, as it is typical in engineering disciplines. To try to overcome these problems we have drafted and tested a new taxonomy that aims at being intuitive, time efficient, and suitable for defining the knowledge and skills required to successfully answer to typical questions and exercises in engineering.

This paper reports our preliminary qualitative and quantitative findings that followed a pilot study.

2. THE PROPOSED TAXONOMY FOR CLASSIFYING EXERCISES

Inspired by the empirical division of engineering knowledge into procedural and conceptual knowledge, the proposed taxonomy assumes the difficulty of a question to be measurable along two dimensions, so that eventually the taxonomy levels are in $\{0, 1, 2, 3\} \times \{0, 1, 2, 3\}$: the ***using*** dimension, dedicated to measuring the difficulty of the procedure needed to compute a correct answer, solve a problem or derive a quantitative result; and the ***explaining*** dimension, dedicated to measuring the difficulty of the reasoning needed to arrive at a correct answer, explain or predict a phenomenon or behaviour, or derive a qualitative result. The taxonomy leverages the ad-hoc concept of *content unit (CU)*, that indicates an atomic unit of teaching-learning content (e.g., "electric potential", "Rouché-Capelli theorem", "Bounded-Input Bounded-Output stability", etc.). Thus we assume that to each question corresponds an opportune set of CUs that indicate which content the question covers.

2.1 The "using" dimension

Exercises that ask explicitly to obtain some specific outputs shall in our taxonomy have a "using" difficulty that is different from zero. Typically those outputs are quantitative and require computations.

Level $u1$ (short for "using level 1") are questions that ask explicitly to obtain some quantitative outputs, and tell explicitly which CUs should be used to compute the outputs, and tell explicitly how to use these CUs, if these can be used in more than one way.

Level $u2$ are questions that ask explicitly to obtain some quantitative outputs, and only hint at which CUs should be used to compute the outputs, and only hint at how to use these CUs.

Level $u3$ are questions that ask explicitly to obtain some quantitative outputs, and neither tell nor hint at which CUs should be used to compute the outputs, and neither tell nor hint at how to use the CUs.

Level $u0$ are questions that do not require to compute a specific output. An example for the "using" dimension is: *Let* $h(t) = \begin{cases} 2 & \textit{for } t \in [0, 2) \\ 0 & \textit{otherwise} \end{cases}$ *be an impulse response of a system. Draw the output of this system if the input is* $u(t) = \begin{cases} 1 & \textit{for } t \in [1, 3) \\ 0 & \textit{otherwise.} \end{cases}$. This exercise is at level $u2$ since it asks explicitly to obtain a quantitative output (thus not $u0$), hints at which CUs should be used to compute the outputs (impulse response, LTI system, signals in the time domain, and thus hinting clearly to convolution), and hints also at how to use the CUs, since convolution as an operation has only one way of being used.

2.2 The "explaining" dimension

Exercises that ask explicitly to obtain some specific qualitative result (such as to explain, motivate or define a content or answer) shall instead have a "explaining" difficulty that is different from zero. We postulate then also 3 distinct levels of "explaining" difficulty

Level $e1$ (short for "explaining level 1") are questions that concentrate on technical aspects and can be answered just through memory recalling operations; such as questions asking the student to define or recall the explicitly mentioned CUs, or recognize the correct keywords or a phrase which define the mentioned CUs. An example of $e1$ questions might be "What is the transfer function of a general discrete-time finite impulse response filter?".

Level $e2$ are questions that concentrate on technical aspects and simultaneously clearly mention or hint both the CUs involved, the technical context, and at least hint at a predescribed or obvious path to reach the solution, and cannot be answered through only memory recalling operations, because they require also performing cognitive / logical connections among the ingredients above to reach an outcome that is explicitly specified in the question, and require logical connections only among the ingredients above. Hence, such questions may ask the student to describe the CUs in own words, provide additional information along with the main points characterizing the CUs, interpret and summarize the CUs, construct a symbolic representation of the CUs, or translate the CUs from one form to another, for example through figures or diagrams. Examples of $e2$ questions might be "Explain how to use a Lyapunov function", or "Describe the role of quality assurance activities in the software process".

Level $e3$ are questions that may include nontechnical content and go beyond the difficulty level $e2$ by presenting at least one of the following features: do not mention explicitly or hint clearly all the ingredients that are needed to answer the question, nor hint at a predescribed or obvious path to get the solution, or require the student to choose from multiple possible nontrivial paths to reach a correct solution, or require constructing on previous knowledge, i.e., performing logical connections with what is beyond what is explicitly mentioned in the exercise, and thus require extrapolating information to correctly predict and/or generalize concepts, consequences and/or phenomena in other contexts and/or outside the subject area. Hence, such questions may ask the student to: • recognize some relationship like similarities, differences, cause-effect relationships, and similar between the ingredients in the question and some non-explicitly mentioned CUs; • identify errors in the presentation or use of some explicitly mentioned CUs; • solve questions that require to apply CUs in some specified situations / contexts but at the same time provide only incomplete information, and thus require the student do logical connections beyond what is explicitly mentioned; • recognize some organizational principles involving the mentioned CUs; • consider some trans-disciplinary aspects of the mentioned CUs; • require to perform analyses, or form opinions, estimates or predictions that necessarily involve what is beyond what is explicitly mentioned in the exercise. Examples of $e3$ questions might be "Comment whether this function may be a valid Lyapunov candidate for this system", "Distinguish the phases of software development process", "Which filter type would you use to solve this problem? Argue your choice." or "Suggest how you would go about validating a password protection system for an application that you have developed".

Of course, questions on level $e0$ also exist, such as questions that require pure computations without any explanation or reasoning.

3. TESTING THE TAXONOMY

After analyzing the processes previously followed by scholars to validate their taxonomies and the scale validation process of Boateng et al. [2018], we started identifying the domain and generating the items for new taxonomy based on existing literature. A content validity assessment followed (the experience and the findings are reported in this contribution); however, validity and reliability testing with a statistically significant sample of faculty and students is planned for the next phase.

According to Boateng et al. [2018] the content validity evaluation is best done through the combination of external expert judges and target-population judges, the sample suggested is from 5 to 7. Therefore, in this phase the sample was recruited via email, in the professional network of the authors and among people that did not participate in the crafting of the taxonomy. Ten people participated voluntarily, covering the role of expert in the subject matter and potential users at the same time. Each meeting was conducted by two project leaders, university teachers in Systems and Control, and a PhD student in Pedagogy. The participants (one female, nine male) were academic members and other professionals, working in Europe (n.7) and Northern America (n.3) in the same scientific area (Systems and Control) and with teaching experience ranging from 0 (meaning at most limited to assisting with preparing exams) to 30 years.

Each meeting followed the same routine, provided the same materials and allocated the same time to each task, so to ensure uniform conditions for each participant. First the participants were asked to read and understand the manual created by the researchers to explain how to use the new taxonomy. As a second task, they were asked to assess the taxonomical level of a set of fifteen exam questions according to the directions written in the manual. The aim of this exercise was to measure whether participants were using the taxonomy consistently, i.e., how dispersed their assessments were, and the provided questions have been previously formulated and tested by the three researchers with expertise in the field. At the end of each meeting a semi structured interview [Trinchero, 2002], was conducted to collect critical issues and suggestions for improvement and explore the dimensions of: **clarity**, especially regarding the lexicon, structure and purpose, so to understand if the taxonomy is described well and unambiguous [Mountrouidou et al., 2019]. **Exhaustiveness**, especially in terms of completeness [Huff et al., 1984] in the sense of being composed by all the domains and categories needed to categorize the difficulty of exercises in an exhaustive way. **Effectiveness**, i.e., if it achieves the established objectives [Alvino et al., 2006], that in this phase of the project is the classification of the difficulty of exercises and the relative labelling. **Relevance** [Devon et al., 2007], especially in terms of usefulness for the purpose of assessing the difficulty of exercises, and usefulness in the assessment process. **Distinctness between categories**, in other words its capability of differentiating questions [Spangler and Kreulen, 2002, p. 667], i.e., having the property for which the defined categories are mutually exclusive [Huff et al., 1984, p. 31]. This means also whether the exercise is “uniquely represented” by difficulty categories [Tett et al., 2000, p. 219].

4. QUANTITATIVE RESULTS ABOUT THE MEASURED ASSESSMENT PATTERNS

Given the small size of the sample (n.10), the results have to be considered as "indications" and as a pilot to design larger experiments, and not as statistical evidence. The main goal of this investigation is to quantify how similarly different experts grade the same exercises.

The results are synthetically reported in Figure 1. For each level of the two scales and each question we highlight in dark green the mode level, i.e., the one that received the largest number of votes, and in light green the level that received the highest number of votes among the levels adjacent to the mode level. Out of 15 questions, only in three questions there are items (Q3 - u dimension, Q7 - e dimension, Q15 - e dimension) where more than 30% of the participants did not converge on one of the "green" levels. This shows a fairly good level of convergence to the same general assessments of the scale. In the average there is a convergence around 80% to the same green "classes" (more precisely 80.5% for the u dimension and 81.0% for the e one).

The convergence to a single class is also fairly good (in average 63.1% converged to the same u level and 51.6% to the same e one), especially if one considers that 10 items (see, e.g., Q1 - u, Q2 - e, Q4 - e, Q8 - e, Q10LU, Q10 - e,Q11 - u,Q11 - e Q13 - e, Q14 - both u and e) have an even distribution between two classes. It must be remarked that, to reduce bias, the questions were extracted from questions used in past automatic-control related courses, and the exercises were not classified prior to this pilot study. As such, the questions were not engineered to clearly belong to a class or another. It is thus very likely that some of the question have levels of cognitive complexity that can be considered intermediate, or between two levels. This realization would have not been possible using "engineered questions" and suggests that, although the taxonomy should remain defined as a discrete scale and users should make the effort to select only one element of the scale, it might be convenient to use continuous numbers in the database indexing tools (similarly to the stars used in online rating systems: users only put discrete value from 1 to 5, that are then aggregated in a continuous number).

Overall the quantitative data show encouraging results and suggest that with some fine tuning on the manual and wording of the taxonomy levels an even higher level of agreement can be achieved. Incidentally, this pilot experiment also showed us the importance of selecting questions/exercise that are very clearly understandable by people coming from different schools and institutional cultures.

5. QUALITATIVE ANALYSIS OF THE USER PERCEPTIONS AND RECOMMENDATIONS

As for the users' perceptions, we focus especially on the content validity, the perceived limitations and the suggestions on how to improve the proposed taxonomy (as described in section 3).

The first considered factor is about the **perceived clarity of the taxonomy**, i.e., the perception of being easy to understand. As for this, all the 10 participants referred that the purpose of the taxonomy is clear. However only 8 perceived its structure as clear, and only 7 perceived its lexicon to be clear. Relative to this, 3 participants identified some words used to describe the dimensions and the related levels of the taxonomy as critical issues. We also note that 2 participants explicitly mentioned that the "explaining" dimension is noticeably less intuitive than the "using" one. We believe that this is likely connected to some issues on the distinctness of the suggested levels.

More precisely, the **perceived distinctness of the levels**, i.e., the clear definition of boundaries among different difficulty levels, is not as accurate as hoped and noticeably non-uniform. E.g., consider Figure 2, listing some specific doubts that may occur while compiling an assessment, and how many times persons felt doubts during their labelling session. Note that the "using" dimension is associated to noticeably less doubts than the "explaining" one, and that "$e2$" vs. "$e3$" seems offering the most fleeting discerning boundary. A shared feeling is that the differentiation between levels seems clear "on paper" (i.e., reading the taxonomy manual and thinking in an abstract way) but then this clarity diminishes when actually trying to apply the taxonomy to the proposed exercises. This calls for adding more examples in the next rewriting of the manual.

The next factor we consider is about the **perceived efficacy of the taxonomy**, i.e., its ability to serve the purpose of classifying different levels of difficulty of different exercises. Every participant agreed on the usefulness of providing some sort of labelling to exercises; 2 of them expressed though concerns that this classification task may intrinsically lead to too coarse results, while 1 expressed doubts about whether classifying difficulties can be performed in a purely objective way at all. Our stance relative to this very important objection is the following: it is unlikely that we will find a taxonomy that removes all subjectivity effects on indexing difficulty levels, however we also believe that proposing a limited tool is better than having no tool at all, especially if we will also be able to characterize the limits of the tool, so as to avoid misuse.

Also, we deal with the **perceived exhaustiveness of the taxonomy**, in the sense of being applicable to all the exercises one may encounter in automatic control related subjects. We note that 7 participants found that there were dimensions missing that should be added to the proposed ones. The most often suggested ones were: *1)* the "time" dimension in the sense of indicating some expected statistics about how much time will be required to an average student knowing how to solve that exercise to actually solve it, and *2)* a "complexity" dimension, in the sense of having a measure that captures how tedious the exercise is. An illuminating example referring to this is

Question	Classes	0	1	2	3
Question 1	LU	4	4	1	1
	LE	1	1	7	1
Question 2	LU	7	1	0	2
	LE	1	1	4	4
Question 3	LU	3	2	1	4
	LE	3	3	4	0
Question 4	LU	2	1	5	2
	LE	3	3	4	0
Question 5	LU	5	2	2	1
	LE	1	3	6	0
Question 6 *(NOTE: one subject did not answer)*	LU	8	1	0	0
	LE	1	8	0	0
Question 7	LU	2	3	5	0
	LE	4	2	2	2
Question 8	LU	7	2	0	1
	LE	2	4	4	0
Question 9	LU	8	2	0	0
	LE	0	2	7	1
Question 10	LU	4	5	0	1
	LE	1	5	4	0
Question 11	LU	3	4	2	1
	LE	1	4	5	0
Question 12	LU	1	8	0	1
	LE	7	2	1	0
Question 13	LU	7	0	1	2
	LE	0	2	4	4
Question 14	LU	8	1	0	1
	LE	0	0	5	5
Question 15	LU	9	0	0	1
	LE	1	3	3	3

Fig. 1. Quantitative summary of the indexing performed by the 10 interviewed teachers about the taxonomy levels of the 15 proposed questions.

"*compute by hand how much 2 times 3 is*" against "*compute by hand how much 2791 times 10123 is*": both questions would be indexable as "using 1, explaining 0", but the vast majority of persons would perceive their difficulty as different. "Time" and "complexity" dimensions appear to be related. Understanding which trade-offs exist relative to adding dimensions (and which ones) is to the best of our knowledge still an open question, and our current research focus.

The next factor we look at is the **perceived relevance of the taxonomy**, in the sense of being a tool that is relevant / useful for teaching purposes. Despite the problems already described above, 7 participants said that already in this form, the proposed taxonomy seems useful for their teaching (especially as a tool for aligning the expectations with the students and colleagues, on top of sharing material within the community). We remark, though, that all the participants were personal acquaintances of the authors, and that the absence of anonymity may have pushed them, also subconsciously, to say kinder words than in other situations.

Finally, as for the **perceived strengths and weaknesses of the taxonomy**, we note that 2 persons explicitly mentioned that already the *existence* of this taxonomy is *per se* a strength. Every participant moreover mentioned at least one particular expected benefit as a strength, the most common ones being: *1)* the possibility of aligning expectations with the various stakeholders on top of exchanging teaching material, *2)* the possibility of checking the consistency of the exams' difficulty levels across the years, and *3)* promoting teachers' reflections on the exercises. As major weaknesses, our perception is that the current taxonomy does not promote enough distinctness between the various levels, and it is insufficiently exhaustive, i.e., it lacks of dimensions to capture the various shades of difficulty of various exercises.

6. CONCLUSIONS

We introduced and piloted a new taxonomy whose purpose is to enable an objective indexing of automatic-control related assessment material. The taxonomy, described in an ad-hoc manual, was used by a sample of 10 people in an indexing exercise accompanied by a semi-structured interview to collect the opinions of the participants. Quantitative and qualitative analyses indicate that *a)* ill-posed questions are evidently associated with higher-than-normal spreads of the indexing, and *b)* some persons are noticeably not consistent with the mean indexing of their peers. The qualitative results indicate that the proposed taxonomy is still incomplete from a dimensional point of view, i.e., it misses some measure of 'how much time' the assessment material requires to the users, and also has issues on being able to classify different levels of difficulties. Future works shall thus focus on reformulating

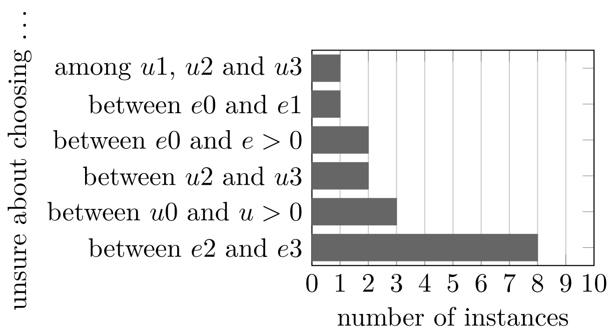

Fig. 2. Number of times doubts were encountered when choosing between two levels during the labeling exercise

the taxonomy so as to address these issues encountered, assessing this new version, and comparing the results so as to understand whether the modifications will have led to improvement. Finally, we believe that a pedagogical research study design could give evidence on how the integration of this taxonomy in the teaching-learning process can support students' learning, giving them a more accurate feedback and assessment on their learning processes and outcomes. In this sense, we are also considering the idea of creating a rubric (i.e., "an assessment tool that explicitly lists the criteria for student work and articulates the levels of quality for each criterion", as well as the scoring strategy used to judge the performance/process [Ragupathi and Lee, 2020, pp.73-74]) in which the levels of quality are identified in accordance with the new taxonomy implemented, to operationalize and facilitate the use of the taxonomy in teaching-learning processes.

REFERENCES

Serena Alvino, Emanuela Busetti, Paola Forcheri, and Maria Ierardi. Networked learning and reusable teaching resources. *Italian Journal of Educational Technology*, 14 (3):30–30, 2006.

Mark Ardis, David Budgen, Gregory W. Hislop, Jeff Offutt, Mark Sebern, and Willem Visser. SE 2014: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering. *Computer*, 48(11):106–109, 2015. ISSN 00189162. doi: 10.1109/MC.2015.345.

Motoei Azuma, François Coallier, and Juan Garbajosa. How to apply the bloom taxonomy to software engineering. In *Eleventh annual international workshop on software technology and engineering practice*, pages 117–122. IEEE, 2003.

John B Biggs and Kevin F Collis. *Evaluating the quality of learning: The SOLO taxonomy (Structure of the Observed Learning Outcome).* Academic Press, 2014.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walquer H Hill, David R Krathwohl, et al. Taxonomy of educational objetives: the classification of educational goals: handbook i: cognitive domain. Technical report, New York, US: D. Mckay, 1956.

B.S. Bloom, D.R. Krathwohl, and B.B. Masia. *Taxonomy of Educational Objectives: The Classification of Educational Goals : Handbook II: Affective Domain.* Number v. 2. MacKay, 1964.

Godfred O Boateng, Torsten B Neilands, Edward A Frongillo, Hugo R Melgar-Quiñonez, and Sera L Young. Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6:149, 2018.

Ricardo Britto and Muhammad Usman. Bloom's taxonomy in software engineering education: A systematic mapping study. In *2015 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE, 2015.

Holli A. Devon, Michelle E. Block, Patricia Moyle-Wright, Diane M. Ernst, Susan J. Hayden, Deborah J. Lazzara, Suzanne M. Savoy, and Elizabeth Kostas-Polston. A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, 39(2):155–164, 2007. ISSN 15276546.

Ursula Fuller, Colin G Johnson, Tuukka Ahoniemi, Diana Cukierman, Isidoro Hernán-Losada, Jana Jackova, Essi Lahtinen, Tracy L Lewis, Donna McGee Thompson, Charles Riedesel, et al. Developing a computer science-specific learning taxonomy. *ACM SIGCSE Bulletin*, 39 (4):152–170, 2007.

John Heywood. *Engineering education: Research and development in curriculum and instruction.* John Wiley and Sons, 2005.

SL Huff, S Rivard, A Grindlay, and IP Suttie. An empirical study of decision support systems. *INFOR: Information Systems and Operational Research*, 22(1):21–39, 1984.

Colin G Johnson and Ursula Fuller. Is bloom's taxonomy appropriate for computer science? In *Proceedings of the 6th Baltic Sea conference on Computing education research: Koli Calling 2006*, pages 120–123, 2006.

Maria Kallia. Assessment in computer science courses: A literature review. *Royal Society*, 2017.

Nurul Naslia Khairuddin and Khairuddin Hashim. Application of bloom's taxonomy in software engineering assessments. In *Proceedings of the 8th WSEAS International Conference on Applied Computer Science*, pages 66–69, 2008.

Susana Masapanta-Carrión and J Ángel Velázquez-Iturbide. A systematic review of the use of bloom's taxonomy in computer science education. In *Proceedings of the 49th acm technical symposium on computer science education*, pages 441–446, 2018.

Tim McMahon and Hemal Thakore. Achieving constructive alignment: Putting outcomes first. *Quality of Higher Education*, 3:10–19, 2006.

Patricia F Mead and Mary M Bennett. Practical framework for bloom's based teaching and assessment of engineering outcomes. In *Education and Training in Optics and Photonics*, page ETB3. Optical Society of America, 2009.

Xenia Mountrouidou, Blaine Billings, and Luis Mejia-Ricart. Not just another internet of things taxonomy: A method for validation of taxonomies. *Internet of Things*, 6:100049, 2019.

Kiruthika Ragupathi and Adrian Lee. Beyond fairness and consistency in grading: The role of rubrics in higher education. In *Diversity and inclusion in global higher education*, pages 73–95. Palgrave Macmillan, Singapore, 2020.

E.J. Simpson. Educational objectives in the psychomotor domain., 60(2). *Behavioral objectives in curriculum development: Selected readings and bibliography*, 60(2): 1–35, 1971.

Scott Spangler and Jeffrey Kreulen. Interactive methods for taxonomy editing and validation. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 665–668, 2002.

Kjell Staffas, Steffi Knorn, Aida Olivia Pereira de Carvalho Guerra, Damiano Varagnolo, and André Teixeira. Using different taxonomies to formulate learning outcomes to innovate engineering curriculum towards pbl: perspectives from engineering educators. In *Educate for the future: PBL, Sustainability and Digitalisation 2020*, pages 310–320. Aalborg Universitetsforlag, 2020.

Alexander Stotsky. Modified solo taxonomy model for constructive alignment in automatic control & signal processing education. Technical report, Chalmers Institute of Technology, 2017.

Robert P Tett, Hal A Guterman, Angela Bleier, and Patrick J Murphy. Development and content validation of a" hyperdimensional" taxonomy of managerial competence. *Human performance*, 13:205–251, 2000.

R Trinchero. *Manuale di ricerca educativa.* Scienze della formazione. Franco Angeli, 2002. ISBN 9788846441584.

MK Watson, J Pelkey, MO Rodgers, and CR Noyes. Exploring student sustainability knowledge using the structure of observed learning outcomes (solo) taxonomy. In *ASEE Annual Conference & Exposition, Indianapolis, Indiana. https://peer. asee. org/20474*, 2014.