Assessment, Evaluation and Measurements 2022 (AEM'22) Welcome from the Mini Track Chair

Stephan Krusche
Department of Computer Science
Technical University of Munich
krusche@in.tum.de

1. Description of the Mini Track

Software engineering deals with complexity and change. It requires the practical application of knowledge. Consequently, its education to the next software engineers that can develop the future systems is an ever-demanding task. Coming generations will deal with even more complex, distributed systems, shorter time-to-market, more complicated quality requirements, build processes, and deployment configurations.

Providing a learning environment that focuses on applying and assessing the actual learning goals based on constructive alignment and in line with research and industry requirements is becoming more and more complex due to the rising number of students in university courses. While individual feedback is essential in learning, it is nearly impossible for instructors of large courses.

Therefore, new strategies in the assessment, evaluation and measurement of students are needed to overcome the weaknesses of existing approaches for automatic grading. Providing consistent feedback while dealing with multiple correction solutions, diverse and heterogeneous student groups and increasingly complex tasks are examples of challenges that belong to this mini-track. In addition, the inclusion of machine learning approaches and the interpretation of assessment data to provide student recommendations while preserving data privacy are research and system development challenges for the next learning tools.

To provide answers to some of these questions, the mini track "Assessment, Evaluation and Measurements" of the special track "Software Engineering Education & Training" was held in conjunction with the 55th Hawaii International Conference on System Science in January 2022. The mini track explores challenges, experiences, approaches, ideas, and new impulses in assessing, evaluating, and measuring students.

In a highly interactive atmosphere, where issues and ideas can be discussed, positioned, and addressed, we sought thought-provoking and highly constructive discussions among a broad audience and presenters to identify promising educational approaches jointly. In addition, we want to try out proposed methods, foster empirical studies, and facilitate collaboration between industry and academia in providing solutions for future learning tools. Accordingly, this mini track accepted contributions focused on, but not limited to, the following topics:

- Assessment approaches for different learning activities in software engineering
- Evaluation of student submissions
- Measurement of student learning activities
- Development of new approaches, workflows, concepts, and tools for auto-grading
- Peer review grading approaches
- The use of teaching assistants and tutors in software engineering education
- Evaluation of soft skills such as presentation and teamwork
- Assessment of creative aspects in software engineering
- Case studies and case examples from university, college, and school courses
- Proposals for and results of empirical studies on assessment, evaluation, and measurements
- Methods and strategies of feedback and grading of student work

2. Review Process

Each paper submitted to the mini track Assessment, Evaluation and Measurements went through a thorough review procedure. At least three experts in software engineering education reviewed each paper following a double-blind process and strictly controlling for conflicts of interest (see Principle 4 in

URI: https://hdl.handle.net/10125/79436
978-0-9981331-5-7

http://www.acm.org/about/se-code#full). We want to thank the following individuals who served as a reviewer (in alphabetical order):

- Bastian Tenbergen, *State University of New York at Oswego (USA)*
- Cecile Peraire, *Carnegie Mellon University (USA)*
- Charles Wallace, *Michigan Technological University (USA)*
- David Kung, *University of Texas, Arlington (USA)*
- Emanuel Grant, *University of North Dakota (USA)*
- Hironori Washizaki, *Waseda University (Japan)*
- Ishtiaque Hussain, *Penn State University (USA)*
- Mark Paulk, *University of Texas at Dallas (USA)*
- Moritz Marutschke, *Ritsumeikan University (Japan)*
- Nancy Mead, *Carnegie Mellon University (USA)*
- Omar Ochoa, *Embry-Riddle University (USA)*
- Richard LeBlanc, *Seattle University (USA)*
- Richard Medina, *University of Hawai'i (USA)*
- Ruth Breu, *University of Innsbruck (Austria)*
- Tim Storer, *University of Glasgow (UK)*
- Timothy Lethbridge, *University of Ottawa (Canada)*

3. Program

The mini track Assessment, Evaluation and Measurements received seven submissions, all rigorously reviewed as outlined above and evaluated on a point system (up to 45 points) with the categories interest, originality, contribution, relevance, theory, methodology, presentation, validity, and references. Papers with an average rating of more than 31 points were accepted, also considering reviewer recommendations. The mini track chair authored one of the submissions and was therefore excluded from the review system. The special track organizers Patrick Letouze and Dan Port took care of the reviews of this submission. Four high-quality papers from the seven submissions were accepted (acceptance rate: 57%).

In [1], Henrik Christensen reports about the experiences of using student-produced screencasts as a medium to explain their solution to advanced design and programming exercises. The screencasts are a viable and relevant alternative to written reports. The author investigated how the assessment of teaching assistants changed based on the new submission format. He concludes that screencast submissions are an essential tool in the teacher's toolbox and provide best practices to gain the full benefits of the approach.

Tenbergen and Daun present the concept of calibrated peer reviews in requirements engineering education [2]. They deal with various compromise solutions for each problem and propose learning by multiple examples to facilitate numerous solution alternatives. Paired with a think pair share model of industry-realistic, project-based milestones, they generated a rich collaborative learning atmosphere. As a result, the calibrated peer reviews significantly improve students' learning outcomes.

Modeling also consists of multiple correct solutions. Krusche proposes a semi-automatic approach for their assessment using a supervised machine learning approach. He aims to increase the fairness and efficiency of grading and improve the provided feedback quality [3]. While tutors manually assess the first submissions, the system learns which elements are correct and appropriate feedback. This approach allows to identify similar elements in subsequent submissions and suggest how to assess them. It promises a high automation rate with few manual adjustments and improved feedback consistency.

Automated grading is now prevalent in many software engineering courses. Clegg, Fraser, and McMinn investigate how different test suites impact grades and the extent to which their observable properties influence these grades [4]. They find a high variation in grades from different test suites, with a standard deviation of 10%. Their findings propose strategies for building test suites that evaluate students' software with consistency based on high coverage, unique and diverse tests, and running tests against artificial faults to determine their quality.

References

[1] Christensen, H., 2022. Using Students' Screencasts as Alternative to Written Submissions. In Proceedings of the 55th Hawaii International Conference on System Sciences.

[2] Tenbergen, B. and Daun, M., 2022. Calibrated Peer Reviews in Requirements Engineering Instruction: Application and Experiences. In Proceedings of the 55th Hawaii International Conference on System Sciences.

[3] Krusche, S., 2022. Semi-Automatic Assessment of Modeling Exercises using Supervised Machine Learning. In Proceedings of the 55th Hawaii International Conference on System Sciences.

[4] Clegg, B., Fraser, B. and McMinn, P., 2022. Diagnosability, Adequacy & Size: How Test Suites Impact Autograding. In Proceedings of the 55th Hawaii International Conference on System Sciences.