eLmL 2024

The Sixteenth International Conference on Mobile, Hybrid, and On-line Learning

ISBN: 978-1-68558-166-4

May 26 - 30, 2024

Barcelona, Spain

eLmL 2024 Editors

José Miguel Jiménez, Polytechnic University of Valencia, Spain

Karsten Boehm, Fachhochschule Kufstein Tirol Bildungs GmbH, Austria

eLmL 2024

Forward

The Sixteenth International Conference on Mobile, Hybrid, and On-line Learning (eLmL 2024), held between May 26th and May 30th, 2024, in Barcelona, Spain, continued a series of events bringing together federated views on mobile learning, hybrid learning, and on-line learning.

eLearning refers to on-line learning delivered over the World Wide Web via the public Internet or the private, corporate intranet. The goal of the eLmL 2024 conference was to provide an overview of technologies, approaches, and trends that are happening right now. The constraints of e-learning are diminishing, and options are increasing as the Web becomes increasingly easy to use and the technology becomes better and less expensive.

eLmL 2024 provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The topics covered aspects related to tools and platforms, on-line learning, mobile learning, and hybrid learning.

We take here the opportunity to warmly thank all the members of the eLmL 2024 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to eLmL 2024. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the eLmL 2024 organizing committee for their help in handling the logistics of this event.

We hope that eLmL 2024 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of mobile, hybrid, and on-line learning.

eLmL 2024 Chairs

eLmL 2024 Steering Committee
Maiga Chang, Athabasca University, Canada
Daniel Amo Filvà, La Salle - Universitat Ramon Llull, Spain
Alexander Mikroyannidis, The Open University, United Kingdom
Rodica Neamtu, Worcester Polytechnic Institute, USA

eLmL 2024 Publicity Chairs
Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

eLmL 2024
Committee

eLmL 2024 Steering Committee

Maiga Chang, Athabasca University, Canada
Daniel Amo Filvà, La Salle - Universitat Ramon Llull, Spain
Alexander Mikroyannidis, The Open University, United Kingdom
Rodica Neamtu, Worcester Polytechnic Institute, USA

eLmL 2024 Publicity Chairs

Sandra Viciano Tudela, Universitat Politecnica de Valencia, Spain
Laura Garcia, Universidad Politécnica de Cartagena, Spain

eLmL 2024 Technical Program Committee

Waqas Ahmed, UniKL Business School (UBIS) - Universiti Kuala Lumpur, Malaysia
Habiba Akter, Universiti Kuala Lumpur (UniKL) Business School, Malaysia
Sikder Tahsin Al Amin, Nabors Industries, USA
Daniel Amo Filvà, La Salle - Universitat Ramon Llull, Spain
Michele Angelaccio, University of Rome “Tor Vergata”, Italy
Luis Anido-Rifon, Universidade de Vigo, Spain
Amara Atif, University of Technology Sydney, Australia
Arlette Audiffred Hinojosa, Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM), Mexico
Elena Bañares Marivela, Universidad Complutense de Madrid, Spain
Verónica Basilotta Gómez-Pablos, Madrid Open University (Udima), Spain
Regina Bäck, KU Eichstätt-Ingolstadt & KSH Munich, Germany
Shreyansh Bhatt, Amazon.com, USA
Krista Bonello Rutter Giappone, University of Malta, Malta / University of Kent, UK
Abinash Borah, Oklahoma State University, Stillwater, USA
Conrad Borchers, Carnegie Mellon University, USA
Tina Boyle Whyte, Marquette University, USA
Egle Butkeviciene, Kaunas University of Technology, Lithuania
Mauro Callejas Cuervo, Universidade Pedagógica y Tecnológica de Colombia, Colombia
Djordje Cantrak, University of Belgrade, Serbia
Laurence Capus, Université Laval, Canada
Cristina M.R. Caridade, Coimbra Polytechnic - ISEC, Portugal
Stefano Caselli, Institute of Digital Games | University of Malta, Malta
Maiga Chang, Athabasca University, Canada
Wanghu Chen, Northwest Normal University, China
Jong Choi, Oak Ridge National Laboratory, USA
Krzysztof J. Cios, Virginia Commonwealth University, USA
Lorraine Cleeton, Northcentral University / Walden University, USA
Aubrey Condor, UC Berkeley, USA
Mike Cosgrave, University College Cork, Ireland
José Paulo Cravino, Universidade de Trás-os-Montes e Alto Douro / CIDTFF, Portugal

Erin Czerwinski, Carnegie Mellon University, USA
Najima Daoudi, École des sciences de l'information, Morocco
Debzani Deb, Winston-Salem State University, USA
Angel P. del Pobil, Jaume I University, Spain
Rafael del Vado Vírseda, Universidad Complutense de Madrid, Spain
Soumyabrata Dev, ADAPT SFI Research Centre, Dublin, Ireland
Federico Diano, University of Naples "Federico II", Italy
Mounim A. El Yacoubi, Telecom SudParis, France
Ramon Fabregat, University of Girona, Spain
Effat Farhana, Vanderbilt University, USA
Xiaojun Fu, Meta Platforms Inc., USA
Ivan Ganchev, University of Limerick, Ireland / University of Plovdiv "Paisii Hilendarski", Bulgaria
Leonardo Garrido, Tecnológico de Monterrey, Mexico
Christos Gatzidis, Bournemouth University, UK
Odeese Ghassa-Khalil, Duquesne University, USA
Henrique Gil, Age.Comm - Instituto Politécnico de Castelo Branco, Portugal
Luca Grilli, Università degli Studi di Foggia, Italy
Alfonso Guarino, University of Foggia, Italy
Sandeep Gupta, Arizona State University, USA
Safa Hamreras, University of 20 Aout 1955, Algeria
Felix Hamza-Lup, Georgia Southern University, USA
Feifei Han, Australian Catholic University, Australia
Hanny Haryanto, Universitas Dian Nuswantoro, Semarang, Indonesia
Joy He-Yueya, Stanford University, USA
Juan Pablo Hernández Ramos, Universidad de Salamanca, Spain
Ángel Hernández-García, Universidad Politécnica de Madrid, Spain
Oriel Herrera, Univ. Católica de Temuco, Chile
Martha Hubertz, University of Central Florida, USA
Francisco Iniesto, Open University, UK / UNED, Spain
Tsunenori Ishioka, The National Center for University Entrance Examinations, Japan
Tomayess Issa, Curtin University, Australia
Katerina Kabassi, Ionian University, Greece
Keiichi Kaneko, Tokyo University of Agriculture and Technology, Japan
Ilias Karasavvidis, University of Thessaly, Volos, Greece
Komalpreet Kaur, Salem State University, USA
Neo Tse Kian, Multimedia University, Malaysia
Carsten Kleiner, University of Applied Sciences&Arts in Hannover, Germany
Samuel Kosolapov, Braude Academic College of Engineering, Karmiel, Israel
Mirjana Matea Kovač, University of Split, Croatia
Laetitia Knight, University of North Texas, Denton, USA / University of Montpellier 3, France
Miroslav Kulich, Czech Technical University in Prague, Czech Republic
Amruth N. Kumar, Ramapo College of New Jersey, USA
Jakub Kuzilek, Humboldt-University of Berlin, Germany
Imane Lamrani, Nikola Motor, USA
Monica Landoni, Università della Svizzera italiana, Switzerland
Cynthia Y. Lester, Perimeter College - Georgia State University, USA
Rory Lewis, University of Colorado at Colorado Springs, USA
Samuel Li, National Center for Atmospheric Research (NCAR) in Boulder, USA

Table of Contents

Wearable Technology and Gaming 1
Qing Li

Bridging Natural Language and Code by Transforming Free-Form Sentences into Sequence of Unambiguous Sentences with Large Language Model 4
Nikita Kiran Yeole and Michael S. Hsiao

Raspberry Pi Controller for Remote Laboratory Hardware Access 11
Adrian Sieber, Nicolas Hentschel, Isabel Heinze, Theo Kaminsky, Michelle Marasas, Rico Beier-Grunwald, Alexander Lampe, Marc Ritter, Christian Roschke, and Matthias Vodel

Digital Peer Grading in Group Learning: Empirical Insights and Best Practices 16
Eduard Klein and Claus Noppeney

Wearable Technology and Gaming

A Study of Teacher Perspectives

Qing Li
Towson University
Learning Design, Technology & School Library Media
Towson University
Towson, USA
e-mail: li@towson.edu

***Abstract*— This study examines teachers' thinking on wearable gaming for educational purposes. The following research questions guide the research: 1) What do teachers believe about wearable gaming in terms of pros, cons, and challenges? 2) What do teachers envision of using wearable gaming for educational purposes? In this case study, data were collected from 31 teachers who enrolled in a graduate course. The analysis of the data showed that conveniences, flexibility, social emotional development, etc. were considered pros, while over reliance on technology, hazards, and inequality, etc. were cons. Interestingly, several other themes were articulated by teachers as both pros and cons. Specific ways to apply wearable gaming for educational purposes were also discussed.**

Keywords- wearable technology; teachers; games; beliefs.

I. INTRODUCTION

The great potential of digital game-based learning is now widely accepted as reflected, for example, in well recognized publications including the Federation of American Scientists' (FAS) report [7] in which games are considered as a powerful tool with great educational potentials [4]. Various studies have demonstrated that games can enhance learners' conceptual understanding (e.g. [9]), motivate students (e.g. [10]), and positively influence players' attitudes (e.g. [1]). On the other hand, wearable technology has increasingly attracted attention from researchers and developers for its power to enhance student learning anywhere and anytime.

Despite the growing interest and increased number of studies in the field, how to best design wearable technology for learning in general, and how wearable technology on Game-Based Learning (GBL) can be optimally used remain to be underexplored [11]. The adaptation of wearable game-based learning in classrooms is scarce due to various reasons. For example, teachers often found challenging to connect wearable games with existing curriculum [3]. The technical skills required to use wearable gaming can be another roadblock for teachers [2]. Using wearable tools can be too complicated for teachers [3].

This study, therefore, aims to bridge this gap by examining teachers' thinking and envisioned use of wearable gaming for educational purposes. Understanding of what teachers concern the most as they consider wearable games for instructional purposes can help us not only better design instructional and training practices for teacher education, but also identify effective approaches for educational wearable game design that are aligned with existing curriculum and meet the needs of the teachers and their students. The following research questions guide the research:

1. What do teachers believe about wearable gaming in terms of pros, cons, and challenges?
2. What are teachers envisioning of using wearable gaming for educational purposes?

II. METHODS

This study was a case study framed in a qualitative, naturalistic research perspective [5]. Aiming to capture teachers' thinking, the focus was on investigating teachers' beliefs about wearable gaming and their envisioned educational use of wearable gaming. Complying with the case study design, this work used a range of data collection approaches to gather detailed information over extended time [5].

The participants were graduate students enrolled in a graduate course involving online learning. A total of 31 students participated which constituted the sample of this study. The course was aimed to provide students with foundational understanding of online education.

The majority of the participants were practicing teachers or formal teachers in k-16 educational institutions with about 20% of them being active or previously worked as trainers in different organizations or businesses. These 31 participants, with about 15% males, were referred to as teachers and pseudonyms were used in this paper. This study was part of a larger research project focused on teacher digital game design experiences. The initial data collection included class observations, assignments completed by teachers, instructor's reflective journal and learners' feedback after class. Other data sources were the teacher created digital artifacts. This paper focuses on participants' reflections, although other data provided information for the context of the study and triangulation of the results.

III. RESULTS

A. Teacher Beliefs

The first research question focused on teacher beliefs about wearable gaming in terms of benefits and drawbacks.

Benefits described by the teachers included: conveniences, flexibility, "[wearable] provide people with convenient, on the go solutions to their everyday dilemmas" (ST). Many argued that wearable gaming would allow highly personalized learning to meet diverse learner needs.

Helping with social emotional development was an advantage cited by many teachers. Some teachers, who initially did not like the idea of wearable technology due to various reasons such as too much screentime, etc. later realized how wearable gaming could help social-emotional development. LT's following comment exemplified this:

- So, my first thought was, give the kids a break. We are inundated with technology as it is... But, as I thought of my students with specific needs during stressful situations, it dawned on me how successful a wearable device would be vs. a timer or adult reminder. It would build personal capacity, independence, and self-awareness way more than an adult reminder or cue would.

Cons and challenges included reliance on technology, hazards, and inequality. The inequality related challenges could be brought by knowledge or language barriers, cost-associated issues as nicely summarized by ST:

- In terms of cons, some include gradual complete reliance on technology, potential hazards, and costs. Potential challenges include inability of some populations to navigate systems due to lack of technological knowledge or language barriers, as well as cost challenges related to obtaining and maintaining the wearable over time (ST).

An interesting observation was that several themes were identified both as pros and cons. The first example was health related topic. From the positive side, teachers articulated how wearable gaming could be used to monitor and thus promote healthy behaviors, but at the same time caution that this might cause "hypervigilance of targeted behaviors" (EF). Additional con included the unknown impact on health from using wearable gaming since "health effects of wearables are unclear" (TB).

The engagement value was the second theme that teachers considered as both beneficial and detrimental. On the one hand, teachers believed wearable gaming would attract students' attention, thus leading to effective learning. On the other hand, concerns were raised about how wearable gaming "could be a distraction" (CP).

The concept of convenience and accessibility was a third topic discussed both as a pro and a con. The teachers repeatedly stated that a benefit of wearable gaming was "being able to be warn provides a sense of convenience" (SM). At the same time, several participants deliberated that a con of wearable gaming was students could suffer from too much exposed to technology, as exemplified by SM: "some parents may not want their children to be exposed to technology consistently."

The last theme that was taken up as both positive and negative related to equity. Wearable gaming was perceived as a tool that could level the playing field because it could allow anyone to access it any time and any place. In contrast, it might create inequality due to various factors such as cost, visually impaired users, social divide, etc.

B. Teacher Envisioned Use

How did teachers envision wearable gaming to be used for educational purposes? The highest number of teachers discussed how they foresaw the use of wearable gaming in helping with daily life skill and functioning. Diverse ideas, ranging from calendar to alarm-type programs, to behavior reminders, were shared as meaningful application of wearable gaming.

One theme that emerged was the integration of wearable gaming with Augmented and Virtual Reality (AR/VR). Examples included "glasses could integrate AR to run scavenger hunts. VR headset could be used for simulations" (TB), or "virtual reality for history (seeing events as they happened)" (MM).

Health related topics, including mental health, were discussed by many participants. How to use wearable gaming to encourage healthy lifestyle both in schools and other workplace settings were mentioned repeatedly.

- Reviewing logs for mood, etc. may allow users to gain insights into times of day or activities that are particularly challenging or health-promoting across their day (EF).
- In a larger workplace setting, challenges between peers would be a great way to use [wearable gaming]. A fitness challenge between coworkers might provide opportunities for involvement in a healthy lifestyle (SF).

A closely related theme identified related to social emotional development. They articulated how wearable gaming could be a valuable tool to help students manage their emotions and improve their social-emotional wellbeing.

- I would love to see a wearable device that supports social emotional well-being by providing breathing techniques with visuals for students to follow. It could encourage who struggle with sharing their emotions to have private opportunity to work on and show those skills (KP).
- In the context of social-emotional development, a built-in reminder to breathe or use a variety of calm down/sensory activities when heart rate increases

due to stress, overstimulation, etc. tailored to specific students (LT).

Content learning, of course, was discussed by some teachers. Examples that were shared by the teachers included:

- Content-based games that are brief & fast paced (e.g. quick math games based on telling time, sight word games, etc.) (SC).
- As a math teacher, I would love to access wearable gaming for educational purposes. If the wearable gaming device allows students to track their speed and time, we could measure a multitude of different scenarios. Using this data...students could solve for equations in relation to their data tables. This type of learning activity would change the way students view algebra concepts (TK).
- In a formal setting I could see using it as a way to incorporate some games into the lesson, such as that one game where the person wears their identity on their head and other people give them clues about it to enhance their social and team building skills (SM).

Equity was another theme identified. Teachers explained how they could use wearable gaming to provide differentiated learning to help diverse learners such as those with special needs.

- In classrooms, students with attention challenges could be quietly prompted to monitor if they are on task or not. I could imagine building reward systems or a game-related component to earn points (EF).
- Wearable gaming could be a great way to seamlessly bring differentiation into a lesson and level the playing field for all students (SR).

Heightening social connection and collaboration to break the brick and mortal boundaries was a salient theme that emerged.

- Students could connect in group activities without having to physically sit next to each other.

IV. CONLUSIONS

The gaming market is still growing with an expected value of US $545.98 billion dollars by 2028, according to the 2021 Fortune Business Insights [8]. Further, gaming is becoming more and more diversified: being played pervasively (e.g. AR games), on new platforms (e.g. VR, mobile games), being played by different groups (e.g. different age levels, both male and females, etc.) [12]. Wearable gaming undoubtedly has its advantages including but not limited to, enabling more flexible user experience through embodied control, and promoting social connections. Yet, wearable game-based learning has little success in education, and we have limited understanding of best practices of using this tool, largely due to its recent emergence. This study addresses the gap in the literature related to wearable gaming and teacher perceptions, adding valuable information to help us understand the value and design considerations of wearables in the context of wearable gaming. Practically, the results of this study are readily understandable by practitioners, which can help guide game designers, developers and educators to best design and use of wearable gaming for educational purposes.

REFERENCES

[1] Y.-J. An and L. Cao, The Effects of Game Design Experience on Teachers' Attitudes and Perceptions regarding the Use of Digital Games in the Classroom. TechTrends, 61(2), pp. 162-170, 2017.

[2] M. Antonioli, C. Blake, and K. Sparks, Augmented reality applications in education. The Journal of technology studies, pp. 96-107, 2014.

[3] M. Bower, D. Sturman, and V. Alvarez, Perceived utility and feasibility of wearable technologies in higher education. Mobile Learning Futures–Sustaining Quality Research and Practice in Mobile Learning, p. 49, 2016.

[4] D. B. Clark, E. E. Tanner-Smith, and S. S. Killingsworth, Digital games, design, and learning: A systematic review and meta-analysis. Review of Educational Research, 86(1), pp. 79-122, 2016.

[5] J. Creswell, Qualitative Inquiry and Research Design; Choosing Among Five Traditions. Sage Publications, 1998.

[6] J. W. Creswell, Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.). Sage, 2014.

[7] Federation of American Scientists, Summit on educational games: Harnessing the power of video games for learning. In: Author Washington, DC, 2006.

[8] F. B. Insights, Gaming Market Worth $545.98 Billion by [2021-2028] Fortune Business Insights, Issue. 2021, https://www.globenewswire.com/news-release/2021/09/22/2301712/0/en/Gaming-Market-Worth-545-98-Billion-by-2021-2028-Fortune-Business-Insights.html

[9] E. Klopfer, S. Osterweil, J. Groff, and J. Haas, The instructional power of digital games, social networking, simulations and how teachers can leverage them, 2009.

[10] Q. Li, Learning Through Digital Game Design and Building in A Participatory Culture: An Enactivist Approach. Peter Lang, 2014.

[11] V. G. Motti, Wearable technologies in education: a design space. Learning and Collaboration Technologies. Ubiquitous and Virtual Environments for Learning and Collaboration: 6th International Conference, LCT 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26–31, 2019, Proceedings, Part II 21.

[12] N. Xi, J. Chen, S. Jabari, and J. Hamari, Wearable gaming technology: A study on the relationships between wearable features and gameful experiences, International Journal of Human-Computer Studies, 181, 103157, 2024.

Bridging Natural Language and Code by Transforming Free-Form Sentences into Sequence of Unambiguous Sentences with Large Language Model

Nikita Kiran Yeole
Computer Science
Virginia Tech, Blacksburg, USA
nikitay@vt.edu

Michael S. Hsiao
Electrical and Computer Engineering
Virginia Tech, Blacksburg, USA
hsiao@vt.edu

***Abstract*—In the realm of natural language programming, translating free-form sentences in natural language into a functional, machine-executable program remains difficult due to the following 4 challenges. First, the inherent ambiguity of natural languages. Second, the high-level verbose nature in user descriptions. Third, the complexity in the sentences and Fourth, the invalid or semantically unclear sentences. Our proposed solution is a Large language model based Artificial Intelligence driven assistant to process free-form sentences and decompose them into sequences of simplified, unambiguous sentences that abide by a set of rules, thereby stripping away the complexities embedded within the original sentences. These resulting sentences are then used to generate the code. We applied the proposed approach to a set of free-form sentences written by middle-school students for describing the logic behind video games. More than 60 percent of the free-form sentences containing these problems were successfully converted to sequences of simple unambiguous object-oriented sentences by our approach.**

Keywords-Natural language programming; decomposition; chain-of-thought reasoning.

I. INTRODUCTION

Natural Language Programming (NLPg) is a concept that attempts to convert instructions/specifications written in free-form natural language into functional program code. NLPg envisions a world in which everyone can program machines without understanding the intricacies of conventional programming languages. While generative Artificial Intelligence (AI) has shown some success in producing code snippets from natural language text, the code that is produced may not adhere to the intent of the input text. When the code does not meet the intent, the user can do one of two things: (1) manually modify the generated code, or (2) re-write the natural language text and try to generate new code. For users who are not experienced programmers, option 1 may not be feasible, since the generated code may contain data structures and/or algorithms that the user is unfamiliar with. Hence, the user is left with the second option. In order to generate functionally correct code, the input must be in a format that the system can process such that common problems with general natural languages are removed. In other words, if the input text is semantically unambiguous, the code generated will more likely adhere to the intent of the input text [1].

An additional benefit is that this helps the user to learn to write unambiguous input text, a necessary skill behind the thought processes in coding. Natural language is increasingly applied in education for personalized AI tutoring and interactive learning, aiding educators in various ways [2] [3] [4]. The ability to instruct a machine in natural language bridges the gap between human thought processes and the digital world, making technology more accessible and intuitive for students.

There are many factors associated with natural language instructions, which makes NLPg extremely challenging [5]. First, the ambiguity in the sentences. Second, the high level verbose descriptions given by humans. Third, complex and compound sentences. Fourth, invalid or erroneous sentences written by humans. We will briefly highlight each of these four areas in the following discussion.

Natural Language (NL) sentences can include ambiguities wherein a single word or phrase may have several interpretations. Consider, for instance, the following English sentence employed in game design:

"When the rabbit touches a rock, it explodes."

Here, the phrase containing the pronoun 'it' creates uncertainty in this sentence. According to one view, the rabbit explodes after touching the rock, whereas the other contends that the rock explodes.

Secondly, the NL instructions can be excessively verbose, especially written by the people who may not know how to program. Consider, for instance, the English sentence employed in game design:

"In a mysterious realm, a lone pointer and some aliens engage in a cosmic dance. When the pointer touches an alien, it changes colors: original to purple, purple to pink. Pink aliens explode."

Here, the sentences provided are verbose with extraneous descriptive words and phrases. Although they adhere to proper English grammar, they deviate from a concise format.

Thirdly, machines typically demand sentences with a clear structure containing a subject, verb, and object. However, complex sentences that sequentially combine multiple events may complicate the parsing of the sentence and prevent a full understanding of the intent of the user. The following sentence illustrates one such example:

"When the carrot turns into a diamond before the carrot touches a fox, the score increases."

Fourthly, when humans provide instructions, there is a chance that they might offer sentences that are invalid, illogical, incomplete or erroneous. In such cases, it becomes difficult for the machine to extract the exact task that needs to be executed. The following is one such example:

"Brick spawns at the bottom. 14 cheese at the top in rows. Ball in the middle. w is up. s is down. brick touches border bounce. ball touches cheese bounces back."

To overcome these challenges, we propose an Artificial Intelligence driven assistant using Large Language Models (LLMs), which will attempt to convert the free-form sentences into sequences of simple sentences, each with a clear subject, verb, and object structure. It promotes a paradigm where instead of the user conforming to the machine, the machine adapts to grasp the user's intent. This assistant streamlines, simplifies, and transforms the NL phrases into directives that machines can easily interpret. The design of the assistant prioritizes rule-driven simplification, methodically translating sentences that eliminate unnecessary elements while retaining the core meaning.

Motivating Example: Consider the following free-form description of a game:

"The rabbit wanders, reversing at borders. The fox wanders, chasing the rabbit when spotting the rabbit. When the rabbit touches the fox, the fox turns into a carrot."

Our goal is to convert the above paragraph to the following simplified sentences.

"There is a rabbit. There is a fox. The rabbit wanders. The fox wanders. If the rabbit reaches a border, it reverses. If the fox sees the rabbit, it chases the rabbit. When the rabbit touches the fox, the fox becomes mutated. When the fox is mutated, it turns into a carrot."

The deconstruction of complex sentences and then rewriting them in basic, simple sentences is the most novel aspect of our strategy. The NL expression frequently combines various thoughts or directives in a single, complex sentence [6]. So, these sentences are decomposed and rewritten in a format that abides by imposed rules. In our approach, the input sentences are parsed, during which the engine identifies key components and breaks them down into their basic elements. By analyzing the relationships between these elements, the system deciphers the user's intention. With this insight, it reconstructs the information into simple sentences that are structured and guided by rules.

The novelty of this paper lies in its specific methodology for simplifying natural language sentences into structured directives through a rule-based system, a departure from traditional semantic parsing and tree-based neural network models which often struggle with the ambiguity and complexity of natural language [5]. We also integrate an educational platform, GameChangineer, to demonstrate the practical application of this approach, showcasing how it facilitates the learning of object-oriented programming concepts by converting these simplified sentences into functional game code.

We applied our approach to process 1000 free-write sentences, out of which 800 sentences contained at least one of the four aforementioned problems, and 200 sentences are non-problematic sentences. The rewritten sentences are then given to an educational platform called GameChangineer [7], [8] that can convert the object oriented English sentences to a functional game [9]. GameChangineer is an AI-Enabled Design and Education Platform which helps students to discover and practice logical reasoning, problem-solving, algorithmic design, critical and computational thinking [7]. Beginners may find Object-Oriented Programming (OOP) to be abstract and challenging to understand due to its emphasis on classes, objects, inheritance, polymorphism, encapsulation, and abstraction. Students can express their thoughts and queries in a way that comes naturally to them when they are able to interact with an educational software through natural language. This reduces the cognitive load associated with learning new, technical syntax and concepts, allowing them to focus more on the underlying principles of OOP. The results showed that more than 60% of the problematic sentences were successfully converted by our approach. The sentences which were successfully converted led to a correct, functional game which adheres to the intent of the user.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 lays out the methodology in our work and Section 4 presents the evaluation of our approach and discusses its implications. Finally, Section 5 concludes the paper.

II. Related Work

A curated list of groundbreaking studies that have had an impact on this field is included in this section.

One approach to addressing these natural language challenges is through semantic parsing, where natural language utterances are encoded and translated into syntactically correct target code snippets using tree-based neural network models [5]. This technique shows promise in generating accurate code snippets from natural language descriptions by focusing on the structural aspects of language to reduce ambiguity and manage complexity. Even sophisticated semantic parsing models, while capable of generating syntactically correct code from natural language inputs, often face difficulties in capturing the user's intent accurately. This is because a single phrase can be interpreted in multiple ways, leading to code that, while technically correct, does not fulfill the intended function [5].

Another sophisticated method involves using execution-based selection processes and Minimum Bayes Risk (MBR) decoding to minimize expected errors in the generated code [10]. This approach selects the most accurate output by considering the execution results of the generated code samples, helping to ensure that the generated code aligns with the intended functionality described in natural language. This approach has its limitations. It requires executing several generated code snippets to determine the best candidate, which can be computationally expensive and inefficient. Furthermore, if the initial pool of generated code contains errors or fails to capture the user's intent accurately, the selection process may still result in sub-optimal code [10].

Deep learning techniques offer significant advancements in understanding and generating code from natural language. By leveraging the encoder-decoder framework, these models can

 ISBN: 978-1-68558-166-4

learn from vast datasets of code to improve the accuracy and relevance of generated code snippets, addressing issues of verbosity and complex sentence structures by focusing on the semantic content of the instructions [11]. Although deep learning has shown promise in understanding and generating code, the models still struggle with sentences that contain multiple actions or intertwined concepts, reflecting a gap in handling real-world complexity [11]. These limitations underline the necessity for a proposed solution that addresses these core issues.

The Transformer model was first presented by Vaswani et al. in their landmark study, "Attention Is All You Need" [12]. In order to deal with ambiguity, the architecture's self-attention mechanism, which is skilled at capturing context, is essential.

Generative pre-trained transformer (GPT)-3 showed its skill in deciphering a wide range of human expressions and offered a solution to unclear or lacking instructions [13]. Despite its outstanding powers, GPT-3 occasionally produces overly detailed or irrelevant answers [13]. GPT-3 also frequently requires particular fine-tuning for certain tasks [13]. BERT's (Bidirectional Encoder Representations from Transformers) pre-training procedure was improved by Liu et al., who published "RoBERTa: A Robustly Optimized BERT Pretraining Approach" [14] [15].

Wei et al.'s study on "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" forms a crucial basis for understanding how Chain of Thought (CoT) in LLMs (Large Language Models) can decompose complex reasoning tasks into a series of simpler, logical steps [16]. The authors demonstrate that CoT prompting significantly improves the ability of LLMs to perform complex reasoning tasks across various domains. We employ CoT not for general reasoning enhancement, but specifically for tackling linguistic challenges in programming, such as verbosity, ambiguities, and complex phrase structures.

We focus on preserving the fundamental semantic meaning of the given instructions while simultaneously addressing the inherent difficulties and limitations of human language. The subtleties of freely written phrases can have a profound impact on the semantic meaning, which is the fundamental core of a communication [17]. Therefore, a major goal in this area should be to transform these statements into more straightforward forms without distorting or losing the original meaning that the user intended. This balance makes sure that, despite the language being more structured or standardized for computational processing, the converted sentences remain true to the message the user intended to convey.

III. Methodology

The foundation of our research is a representative dataset, which was used as the LLM's main input. The data included 1000 student-written free-form sentences as game descriptions. 800 of these sentences have been identified as potentially problematic and 200 sentences have been identified as non-problematic. These descriptions offered a variety of linguistic patterns and semantic complexities. The game descriptions were diverse, varied in their lengths, and offered a number of difficulties. These sentences showed some ambiguity because they frequently contained intricate structures and relationships that were not always clear. This dataset was also chosen to evaluate the LLM's capacity to comprehend and translate the ambiguous and complex texts into more rule-based, simplified formats.

We used the GPT-3.5 Turbo, a powerful language model created by OpenAI, for the purposes of this study. We made this choice after carefully comparing the performance of GPT-3.5 Turbo and GPT-4, two recent revisions of OpenAI's generative models. Although GPT-4 is a more recent model and is anticipated to offer higher capabilities in many contexts [18], GPT-3.5 Turbo showed improved sentence construction in the most basic form and coherence for the particular prompt utilized in this research. This underscored the need of selecting a model that is tailored to the precise specifications of the work at hand as opposed to just selecting the most recent version. This model was deployed by means of direct integration with the OpenAI API, which allowed us to operate the model locally in our computational environment. Python was selected as our primary programming language because of its extensive libraries for data manipulation and its seamless integration with the OpenAI API.

The model's temperature was set to zero. The choice was made to guarantee deterministic performance from the model.

The top_p parameter was set to 1. This implies that at each stage of the generation process, the model will only take into account the tokens that are the most likely.

It should be emphasized that these combinations signify that we used the model outside of its intended parameters. We purposefully restricted the model to create consistent and repeatable results customized to our needs rather than utilizing its potential for creative and varied outputs. These settings came in helpful in situations where consistency and predictability were crucial.

Our method employed a split strategy that made use of both user prompts and system prompts. The user prompt constitutes the primary interaction point with the user. It is necessary to convert these user-provided free-form sentences into a (sequence of) more simplified structure. The model must understand these inputs robustly due to the inherent variation in how users phrase their queries or utterances. Free-form phrases can be anything from simple sentences to more complex thoughts or assertions, and the challenge lies in distilling the essence of what the user wants to communicate and converting it into a form that the model can process efficiently.

The system prompt serves primarily as a tool to direct the model towards a specific context or mode of operation. We directed the model's potential and ensured that we receive the desired output by creating a structured system prompt. It encompasses a chain-of-thought reasoning via (1) Question Answering, (2) Sentence Reframing, (3) Sentence Decomposition. Figure 1 shows the process flow with an example prompt for each step.

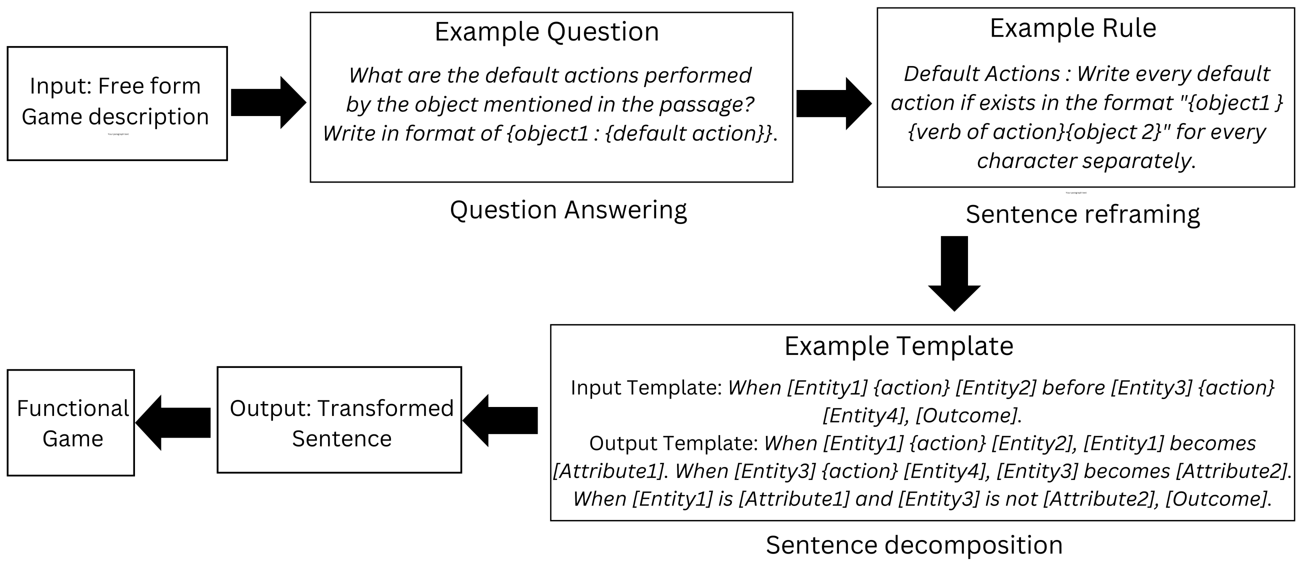


Figure 1. Process flow with example prompt for each step.

A series of iterative tests and comparisons with additional approaches, such as few-shot learning and model fine-tuning, revealed that the suggested strategy performed better overall, especially with unrestricted sentence structures.

Let us consider an input text:
The apricot slows down at border. The rabbit turns into a diamond when hitting a carrot.

Here is a step-by-step trace through the outlined process using the provided input sentence.

1) Question Answering (QA): The QA component extracts crucial information from the input sentence by asking questions and taking the output in a specific format. It identifies the objects (apricots, rabbits, borders, diamonds), the default actions (apricots and rabbits move), and the conditional actions (speed decrease for apricots, transformation for rabbits).
2) Sentence Re-framing: Using the information from the above QA, the sentences are then re-framed according to a set of predefined rules that reflect the original free-form sentences. The main goal here is to use a specified set of rules to reconstruct the sentences in a paragraph which are in their basic form in the format subject-verb-object. For example, stating the conditional actions of various objects: when apricots touch a border, their speed decreases, and when rabbits touch a carrot, they turn into diamonds.
Re-framed sentence: If the apricot touches a border, the speed of the apricot decreases. If the rabbit touches a carrot, the rabbit turns into a diamond.
3) Sentence Decomposition: Next, the Sentence Decomposition step would break down complex sentences into simpler, object-oriented structures. The input would be analyzed to discern patterns of object interactions, such as the apricot's speed change upon touching a border, and the rabbit's transformation upon touching a carrot. An intermediate attribute ”mutated” is added while decomposing the sentence resulting in the following sequence of unambiguous sentences [19].
Decomposed sentence (Final Output): If the apricot touches a border, the speed of the apricot decreases. When the rabbit touches a carrot, the rabbit becomes mutated. When the rabbit is mutated, it turns into a diamond.

To sum up our methodology, it offers a comprehensive, structured, and systematic approach to interpret and process natural language text with a high degree of precision and consistency, enabling the user to more accurately describe their intent. Our innovation lies in the strategic application of existing LLM capabilities through a series of system prompts that guide the model to produce outputs in line with specific, predefined rules. This ensures that the transformations maintain the core meaning of the original sentences while stripping away unnecessary complexities, making the text more suitable for generating executable code.

Few-shot learning was initially considered due to its prowess in addressing edge cases with limited data. However, given the vast array of edge cases, rules, and potential issues to address in this domain, few-shot learning proved insufficient. The model would occasionally produce out-of-bound prompts leading to sub-optimal performance. In contrast, our proposed approach, which integrates QA, reframing, and sentence decomposition exhibits robustness against diverse sentence structures, making it an ideal choice for our purpose.

IV. Evaluation and Discussion

This section evaluates the performance of the proposed AI-driven assistant in processing 1000 free-form sentences categorized into five types: (1) Grammar/typos, (2) Ambiguous, (3) Unrealizable actions, (4) Overly complex/descriptive, and (5) Non-problematic sentences. Sentences containing grammatical or typographical errors fall under the first category, ”Grammar or Typos” that could cause misinterpretations or inaccurate code translations. The second category, ”Ambiguity” refers to statements that have ambiguous references or meanings. Examples of this type of sentence include ”It chases it”, where pronouns make it difficult to determine exact entities and actions. The third category, ”Unrealizable Actions”, consists of sentences that describe actions not feasibly translatable into programming logic, exemplified by phrases like ”It jumps to heaven”. Sentences falling into the ”Overly Complex or Descriptive” category are weighed down with too many information or complex structures, which makes it difficult to translate them into concise, executable computer commands. Each of these categories represents a unique facet of the complexity inherent in translating natural language into machine-executable code. The final ”Non-problematic sentences” category refers to the sentences which are successfully translatable by the GameChangineer platform into executable code [7] [8] [9]. These sentences are unambiguous and in object oriented structure.

There are several reasons why the final category of ”Non-problematic sentences” is included. It serves primarily as a benchmark, providing a point of comparison to assess the efficiency and precision of the AI-powered assistant while processing and interpreting texts that do not present inherent challenges. Furthermore, this category aids in determining whether and how Language Models (LMs) intervention may unintentionally add errors into previously error-free sentences. This will help in evaluating the preservation of sentence integrity after processing and is essential for preserving the overall quality and validity of the research.

The above categorization is based on the platform’s algorithms that use symbolic AI to detect grammatical errors, ambiguity, complexity, and unrealizable actions in sentences, indicating potential issues for translating these into executable code. The platform automatically logs the problematic sentences. All logged erroneous sentences are analyzed in this paper.

We discuss the effectiveness of the assistant in identifying and rectifying these issues, thereby enabling accurate translation into executable code. These sentences were written by middle school students with different degrees of experience in both natural language expression and game design when they were first created as parts of game descriptions. This diversity guarantees a wide range of linguistic difficulties, reflective of the intricacies typically seen in natural language programming.

These middle school students received a basic introduction to writing a few simple games with the GameChangineer platform. A small percentage of the students have prior programming experiance. However, a vast majority of the students have never programmed before. Participants were given the following instructions to create their game plan: ”Write a game plan for creating a game utilizing the available characters.”

To ensure the accuracy and feasibility of the translated sentences produced by the LLM, they were given as an input into the GameChangineer platform [7]. This platform provides a score for each sentence that measures the compatibility with the platform’s expected input format [7] [8] [9]. Although some complex sentences can already be decomposed into a sequence of sentences by the GameChangineer platform, it cannot process all the nuances in natural language. We note that all the original problematic sentences were not accepted by the GameChangineer platform.

After the original input sentences were re-written by the LLM using our proposed approach, these new sentences underwent the validation process. Whenever the rewritten sentence(s) are understood with more than 90% certainty by the GameChangineer platform, the conversion will be regarded to have been translated correctly; on the other hand, when it falls below this mark, the output program generated may contain errors. The output program is generated by the GameChangineer Platform. The accuracy and relevance of the LLM-generated results were also assessed manually to ensure the translations effectively communicated the intended meaning. This dual evaluation provides a comprehensive measure of the AI assistant’s efficacy in translating complex natural language into machine-executable code by combining automated accuracy assessment with manual semantic verification.

Table 1 presents the results of the sentence categorization from the data-set, highlighting the success rate for each category. The table is divided into three main columns: Sentence Category, Number of Sentences, and Success Rate. These categories include Grammar/Typos, Ambiguous, Unrealizable Actions, Overly Complex/Descriptive, and Non-problematic. Note that when all 5 categories are considered, the success rate was more than 68%. However, if we consider only those first 4 categories (excluding the Non-Problematic category), the success rate by our approach is more than 60%.

The category of Unrealizable Actions encompasses game descriptions featuring actions that are either illogical or infeasible within the game context. This category was the least represented in student written game descriptions, accounting for its minimal proportion. All erroneous sentences are logged by the platform and the distribution of categories reflect the distribution of error types. The presence of actions that cannot be executed within the game complicates the task of rewriting such sentences, contributing to the lowest success rate observed for this category.

A sample of successful and unsuccessful transformations are listed in Table 2. In the following, we will explain a few specific cases.

Let us first examine a successful transformation:

Input Sentence: ”Each carrot you eat it gives you 1o speed of

TABLE I. SENTENCE CATEGORIZATION RESULTS

Category	**Sentence proportion**	**Success Rate**
Grammar/Typing mistakes	22%	75.6%
Ambiguous Sentences	18%	57.7%
Unrealizable Actions	11.6%	50%
Overly Complex/Descriptive	28.4%	59%
Non-problematic sentences	20%	94%
Total Success Rate	100%	68.4%

TABLE II. INPUT AND OUTPUT SENTENCES

Input Sentence (Free-form)	**Output Sentence**	**Successful?**
fox explodes rabbit when it is the rabbit is touched by it.	If the rabbit is touched by the fox, the fox explodes.	Successful
When the fox touches the carrot before the rabbit, the fox dies.	When a fox touches a carrot, the fox becomes happy. When a rabbit touch the carrot, the rabbit becomes happy. When the fox is happy and the rabbit is not happy, the fox dies	Successful
if the fox sees the rabbit and the fox chases it and catches it the rabbit explodes.	When the fox sees the rabbit, it chases the rabbit. When the fox catches the rabbit, the rabbit explodes.	Successful
Rabbit move with wsad.	When the W key is pressed, the rabbit moves up. When the S key is pressed, the rabbit moves down. When the A key is pressed, the rabbit moves left. When the D key is pressed, the rabbit moves right.	Successful
The kitten must touches all the cakes to win without having the fox touch the kitten.	When the kitten touches all the cakes, the player wins the game. If the fox touches the kitten, the player loses the game.	Successful
When cake is touched by rabbit the speed of the rabbit is 0.5 for 3 seconds and cake disappears.	When the rabbit touches the cake, the rabbit's speed becomes 0.5 for 3 seconds. When the rabbit touches the cake, the cake disappears.	Successful
When the ball touches the bottom border it gains 1 point.	When the ball touches the bottom border, the score adds 1.	Successful
the pointer shoot the ball with the spacebar.	The pointer shoots the ball when the spacebar is pressed.	Successful
The rabbit dies when it touches the diamond for 5 seconds.	When the rabbit touches the diamond for 5 seconds, the rabbit becomes dead. When the rabbit is dead, it dies.	Successful
The rabbit must flee from the fox.	If the rabbit sees the fox, the rabbit flees.	Successful
the sheep is trying to eat all the broccoli.	If the sheep sees broccoli, the sheep eats broccoli.	Successful
When a tiger is shot for the 2 times, it disappears.	If a tiger is shot twice, it becomes mutated. When a tiger is mutated, it disappears.	Successful
if you do not touch the apricot within 3 minutes game is over.	If the apricot is not touched within 3 minutes, the game is over.	Unsuccessful
If the donut is blue and W is pressed, the donut moves up	There is a donut. When the W key is pressed, the donut moves up 10. If the donut is blue, then the color of the donut changes.	Unsuccessful
fox explode 20 seconds after start.	There is a fox. If the start is 20 seconds, the fox explodes.	Unsuccessful
The panda kills foxes when it reaches a score of greater than or equal to 14.	If the score is greater than or equal to 14, the score adds 1.	Unsuccessful
There are 3 snakes going side to side at 3 pixels per frame.	The snakes move side to side at 3 pixels per frame.	Unsuccessful

the rabbit."
Transformed Sentence: "When the player eats a carrot, the speed of the rabbit increases by 10."

The rewritten sentence establishes a clear subject "the player" and action "eats a carrot" by explicitly stating "When the player eats a carrot" as the condition. The usage of "it" in the original sentence lacks a clear antecedent, which could cause parsing issues. This ambiguity is eliminated in the rewritten sentence by defining the action and its result clearly without the use of potentially confusing pronouns. The phrase "the speed of the rabbit increases by 10" explicitly describes the effect of the action in a structured way that is easy for a machine to parse. This approach makes it easy to transform into a computational operation because it explicitly states the change ("increases by 10") and the attribute being impacted ("the speed of the rabbit").

Let us look at a few examples that show how the AI-driven assistant can recognize and fix typos and grammatical problems in sentences.
User Input: "When Dis pressed the bird moves right."
Transformed Sentence: "When the D key is pressed, the bird moves right."

The original sentence contained a typo ("Dis" instead of "D") and lacked a verb, making it unclear. The AI model successfully recognized the typo, corrected it to "D", and inserted the necessary word "key" to complete the sentence.
User Input: "When the rabbit is silver, the speed of the rabbit."
Transformed Sentence: "If the rabbit is silver, the speed of the rabbit increases."

The user input was incomplete, with no clear action described for the rabbit. The AI assistant inferred the most logical action "speed increase" based on the context, making the instruction actionable for programming.

Consider one unsuccessful transformation which could not adhere to the intent of the user.
User Input: "When the rabbit is yellow, the rabbit touches the fox, the fox explodes."
Transformed Sentences: "If the rabbit is yellow, the rabbit touches the fox. If the rabbit touches the fox, the fox explodes."
Intended Output: " If the rabbit is yellow and the rabbit touches the fox, the fox explodes."

The input is incorrectly divided into two conditional statements by the original transformation. This method falsely implies that the fox's explosion is a two-step process that depends on the requirements being satisfied sequentially as opposed to simultaneously. The intended output, on the other hand, combines the two conditions into a single compound condition meaning that the fox will explode if both conditions are met simultaneously and directly. This showed that the input sentence is ambiguous and the AI-assistant could not successfully transform the sentence.

Let us look at an unsuccessful example in the fifth category, Non-problematic sentences.
User Input: "When a ball sees the rock, the ball flees from the rock."
Transformed sentences: "When the ball sees the rock, the ball becomes scared and flees from the rock."

The transformed sentence is considered unsuccessful here, primarily due to the addition of an unwanted attribute ”scared” to the output sentence. This is an example where the LLM hallucinated leading to add an extra and unnecessary attribute [20]. Such hallucinations can significantly impact the utility and accuracy of LLMs, especially in applications requiring strict adherence to input data without the addition of interpretative or speculative elements. LLMs occasionally ”hallucinate,” or provide missing information [20]. We found that unsuccessful conversions due to hallucination account for 6% of Non-problematic sentences. For the problematic sentences in the other four categories, hallucination is responsible for about 12% of the unsuccessful transformations.

We did not compare our results with LLM based code generation platforms such as Copilot [21] because our goal is to rewrite erroneous sentences so that they become clear and unambiguous. On the other hand, while Copilot may be able to generate code on an erroneous sentence, it generates the code by its own interpretation arbitrarily. In addition, GameChangineer can process hundreds of sentences at a time, but the user must interface Copilot differently by feeding a few sentences at a time.

V. Conclusion and Future Work

This paper presents a method of converting free-form natural language sentences into a sequence of unambiguous, simplified sentences that can subsequently be translated into machine-executable code. The utilization of LLMs has shown promise in addressing the inherent difficulties brought about by verbosity, ambiguities, complexity, and possible errors. Our approach, which combines aspects of Question Answering, Sentence Reframing, and Sentence Decomposition has demonstrated a notable capacity to handle a wide variety of linguistic patterns and semantic complexities. More than 68% of the 1000 problematic and non-problematic sentences were correctly converted by the proposed method.

There are areas for improvement, particularly in understanding complex conditional relationships and refining the LLM methodologies, aiming to reduce the incidence of hallucinations. The results highlight the inherent challenges of processing natural language, particularly in dealing with the nuances of human language. Additionally, they draw attention to how AI-powered systems have the potential to greatly enhance our comprehension and interpretation of words with unclear structures, which is an important area of study in the field of natural language programming.

Acknowledgement

This research is supported in part by NSF grant 2101021.

References

[1] C. Yang, Y. Liu, and C. Yin, “Recent advances in intelligent source code generation: A survey on natural language based studies,” *Entropy*, vol. 23, no. 9, p. 1174, 2021.

[2] D. Baidoo-Anu and L. Owusu Ansah, “Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of chatgpt in promoting teaching and learning,” *SSRN Electronic Journal*, January 2023, published by Elsevier BV, retrieved: March, 2024, Available at SSRN: https://ssrn.com/abstract=4337484 or http://dx.doi.org/10.2139/ssrn.4337484.

[3] T. P. Tate, S. Doroudi, D. Ritchie, Y. Xu, and M. W. Uci, “Educational research and AI-generated writing: Confronting the coming tsunami,” January 2023, published by Center for Open Science, retrieved: April, 2024. [Online]. Available: https://doi.org/10.35542/osf.io/4mec3

[4] D. Mogil et al., “Generating diverse code explanations using the GPT-3 large language model,” in *ICER ’22: Proceedings of the 2022 ACM Conference on International Computing Education Research*. Association for Computing Machinery, 08 2022, pp. 37–39.

[5] F. F. Xu, B. Vasilescu, and G. Neubig, “In-IDE code generation from natural language: Promise and challenges,” *ACM Trans. Softw. Eng. Methodol.*, vol. 31, no. 2, mar 2022, retrieved: April, 2024. [Online]. Available: https://doi.org/10.1145/3487569

[6] C. Niklaus, “From complex sentences to a formal semantic representation using syntactic text simplification and open information extraction,” Ph.D. dissertation, 03 2022, retrieved: April, 2024. [Online]. Available: https://opus4.kobv.de/opus4-uni-passau/frontdoor/index/index/docId/1054

[7] M. S. Hsiao, “Automated program synthesis from object-oriented natural language for computer games,” in *Proceedings of the Controlled Natural Language Conference*, August 2018.

[8] ——, “Multi-phase context vectors for generating feedback for natural-language based programming,” in *Controlled Natural Language*, September 2021.

[9] ——, “Automated program synthesis from natural language for domain specific computing applications,” Patent 10 843 080, November, 2020.

[10] F. Shi, D. Fried, M. Ghazvininejad, L. Zettlemoyer, and S. I. Wang, “Natural language to code translation with execution,” 2022, arXiv.

[11] T. H. M. Le, H. Chen, and M. A. Babar, “Deep learning for source code modeling and generation: Models, applications, and challenges,” *ACM Computing Surveys*, vol. 53, no. 3, p. 1–38, Jun. 2020, retrieved: April, 2024. [Online]. Available: http://dx.doi.org/10.1145/3383458

[12] A. Vaswani et al., “Attention is all you need,” in *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017, pp. 5998–6008, retrieved: April, 2024. [Online]. Available: https://arxiv.org/abs/1706.03762

[13] M. L. Zong and B. Krishnamachari, “A survey on GPT-3,” 2022, arXiv, retrieved: April, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:254221221

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *North American Chapter of the Association for Computational Linguistics*, 2019, retrieved: April, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:52967399

[15] Y. Liu et al., “Roberta: A robustly optimized BERT pretraining approach,” 2019, CoRR, abs/1907.11692, retrieved: April, 2024. [Online]. Available: http://arxiv.org/abs/1907.11692

[16] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 24 824–24 837.

[17] J. Akanya and C. G. Omachonu, “Meaning and semantic roles of words in context,” *International Journal of English Language and Linguistics Research (IJELLR)*, vol. 7, pp. 1–9, 03 2019.

[18] M. Rosoł, J. S. Gasior, J. Łaba, K. Korzeniewski, and M. Młyńczakl, “Evaluation of the performance of GPT-3.5 and GPT-4 on the polish medical final examination,” *Scientific Reports*, vol. 13, no. 1, p. 20512, 2023, retrieved: April, 2024. [Online]. Available: https://doi.org/10.1038/s41598-023-46995-z

[19] N. Wies, Y. Levine, and A. Shashua, “Sub-task decomposition enables learning in sequence to sequence tasks,” 2023, arXiv.

[20] Z. Ji et al., “Survey of hallucination in natural language generation,” *ACM Computing Surveys*, vol. 55, no. 12, p. 1–38, Mar. 2023, retrieved: April, 2024. [Online]. Available: http://dx.doi.org/10.1145/3571730

[21] GitHub, “About github copilot,” 2024, retrieved: April, 2024. [Online]. Available: https://docs.github.com/en/copilot/about-github-copilot

Raspberry Pi Controller for Remote Laboratory Hardware Access

Adrian Sieber, Nicolas Hentschel, Isabel Heinze, Theo Kaminsky, Michelle Marasas,
Rico Beier-Grunwald, Alexander Lampe, Marc Ritter, Christian Roschke, Matthias Vodel
Hochschule Mittweida - University of Applied Sciences
Technikumplatz 16, 09648 Mittweida, Germany
Email: {esieber, nhentsch, iheinze, tkaminsk, mmarasas, beier1, lampe, ritter, roschke, vodel}@hs-mittweida.de

***Abstract*—Throughout the pandemic, institutional education switched from face-to-face lectures and practical application to remote learning facilitated by learning management platforms and remote laboratory systems. While conventional lecture formats were replaceable, especially in engineering laboratory work was limited by available remote laboratory systems, the possibility to integrate them as well as their range of functions and financial attractiveness. We introduce RaspCon, an easily implementable Raspberry Pi controlled web-application based on open-source software, as an extension to the LabCon system, allowing not only for remote access to real laboratory hardware but also the reconfiguration of its interconnections.**

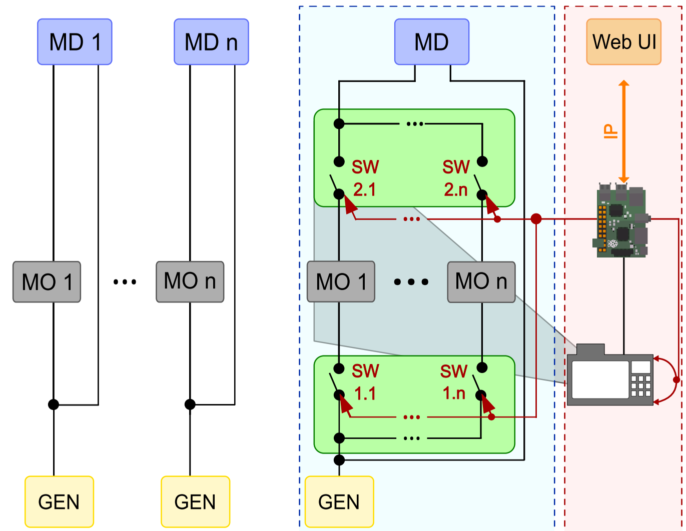
***Keywords*-digital learning; e-learning; raspberry pi; remote hardware access; web application.**

I. Introduction

In response to the challenges posed by the COVID pandemic, the landscape of education underwent a rapid transformation from traditional face-to-face instruction to remote learning modalities. This shift necessitated the utilization of digital platforms, such as video conferencing software like Zoom and Microsoft Teams, alongside Learning Management Systems (LMS), such as Moodle, Opal, and Adobe Captivate. However, work-based learning activities, like laboratory work, faced considerable hurdles in adapting to remote environments. Remote Lab Systems (RLS) that allow and manage access to professional laboratory equipment are offered by several companies [1][2], but these often incur substantial costs. Moreover, adapting them to support existing laboratory practicals or integrating them into standard LMS platforms, including user authorisation, has proven to be partly challenging.

In recent years, a proliferation of cost-effective solutions has emerged, leveraging open-source software alongside readily available, affordable hardware like Arduino or Raspberry Pi boards. These solutions have been tailored to cater to a diverse array of basic experimental applications, as exemplified in [3] [4][5] and references therein. Additionally, remote laboratories featuring commercial-grade industrial hardware have been developed. For instance, Grodotzki et al. introduced a remote lab designed for conducting automation and control experiments with actual robots [6]. Similarly, García et al. described a tool for managing PCs utilized in student experiments [7].

Achilles et al. developed LabCon [8], a web-based RLS integrated with LMS functionality based on open-source software that enables authorised users to gain access to laboratory hardware. Employing the LabCon system for remote experimentation involving multiple Measurement Objects (MOs)



(a) Initial setup for n MOs requiring multiple GENs and MDs.

(b) New setup dynamically connecting a single GEN and MD to n MOs using electronic switch matrices (light blue box) controlled by RaspCon (light red box)

Figure 1. Comparison of initial and new LabCon hardware setup.

typically requires the hard wiring of one signal source or GENerator (GEN) and one Measurement Device (MD) for each measurement object, as depicted in Figure 1a. Any adjustments to the configuration demand manual intervention, compelling the physical presence of an operator within the laboratory premises. To enhance flexibility, eliminate the necessity for manual hardware reconfiguration, and reduce the number of instruments needed, digitally controllable electronic hardware switch matrices have been developed [9]. As highlighted in light blue in Figure 1b, these matrices facilitate the routing of the generator signal to the input of any of the multiple measurement objects and connecting any measurement object's output to an input channel of the measurement device.

To facilitate remote operation of this enhanced hardware setup, an open-source software based application named RaspCon was developed, as shown within the light red box in Figure 1b. RaspCon enables intuitive web-based remote control of the electronic hardware switches and integrates with a rotatable camera to provide a live stream from the laboratory for visual inspection. In this paper, we will elucidate the motivation behind and intricacies of RaspCon, detailing its integration with the initial LabCon system to form the extended LabCon framework. Section II introduces the given requirements and outlines the derived architecture of the RaspCon system. Section III describes the hardware components of RaspCon and their specifications. The controller software and Web UI are presented in Sections IV and V, respectively. Results from a user evaluation are presented in Section VI. Finally, Section VII provides a summary and outlook.

II. Requirements and System Architecture

As previously outlined, RaspCon is designed to allow for intuitive remote control of electronic switch matrices via a Web-based User Interface (Web UI), provide live camera streams from the laboratory, and seamlessly integrate into the existing LabCon framework. Consequently, the major components of RaspCon must support the following key features:

- Controller Hardware:
  - Host an IP-accessible web application.
  - Provide a sufficient number of programmable output pins to control external hardware.
  - Include a high-bandwidth camera connector.
  - Be scalable and cost-efficient.
- Controller Software:
  - Generate and apply digital control signals for hardware switches.
  - Manage preset digital experimental setups.
  - Support the streaming of live video signals. Offer a robust web server.
  - Be seamlessly integrable into the existing LabCon system to leverage its user access functions (see [8]).
  - Use open-source software components.
- Web UI:
  - Support intuitive usage for ease of operation.
  - Be resilient to misuse to ensure system integrity.
  - Provide a positive user experience.

The architectural overview in Figure 2 illustrates the integration of the current LabCon system with the new laboratory hardware setup and the RaspCon controller.

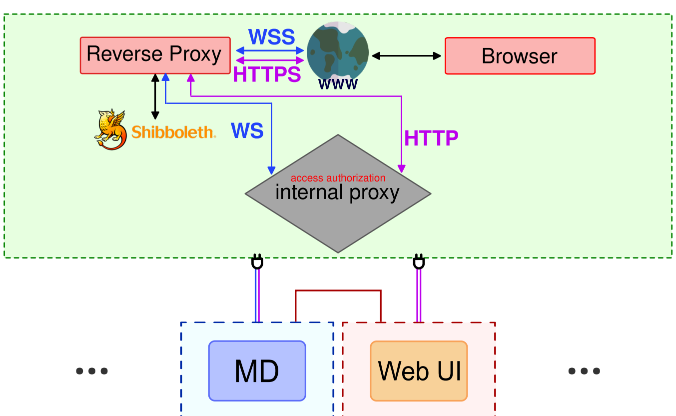


Figure 2. LabCon System Architecture: current LabCon system (light green box), LabCon laboratory hardware setup with electronic switch matrices (light blue box) and RaspCon controller (light red box).

III. Controller Hardware and Camera Module

A careful selection process was undertaken to ensure that the hardware controller would meet all requirements, especially in communicating with the electronic switch matrix, hosting the web application, and supporting video streaming.

Several single-board computers were evaluated for their suitability as controller hardware, including the Arduino, Raspberry Pi, and Banana Pi. Ultimately, the Raspberry Pi Model 4B was chosen for its superior performance and versatility. This decision was driven primarily by two key factors: Firstly, its robust processing capabilities that enable the hosting of a full Linux operating system like Raspberry Pi OS, thus providing flexibility beyond embedded programming languages when choosing the software stack. Secondly, its rich range of hardware interfaces and General-Purpose Input-Output (GPIO) pins necessary for controlling both current and potential laboratory hardware, servo motors and camera. Furthermore, the Raspberry Pi's widespread adoption ensures an extensive array of hardware extensions and libraries, reducing the need for custom hardware extension development [5]. Additionally, the Raspberry Pi's accessibility and long-term support were pivotal considerations.

In addition to the Raspberry Pi single board computer, a Raspberry Pi 2 camera module was selected to remotely monitor the laboratory experiments. Should there be a need to adjust the camera specifications in the future, the system can seamlessly accommodate any device compatible with Video4Linux2 (V4L2). The camera module itself is mounted on a pan-tilt HAT, allowing for dynamic adjustment of the viewing area through rotation and tilting. This functionality is achieved using servo motors controlled via pulse-width modulated signals.

IV. Controller Software

An overview of the RaspCon software stack is depicted in Figure 3. In our pursuit of an efficient and developer-friendly solution, we opted for the fullstack framework SvelteKitas the foundation of our application. SvelteKit, built upon the innovative Svelte framework and running on top of the Javascript runtime environment Node.js, offers a multitude of advantages that align closely with our project requirements. One of the key

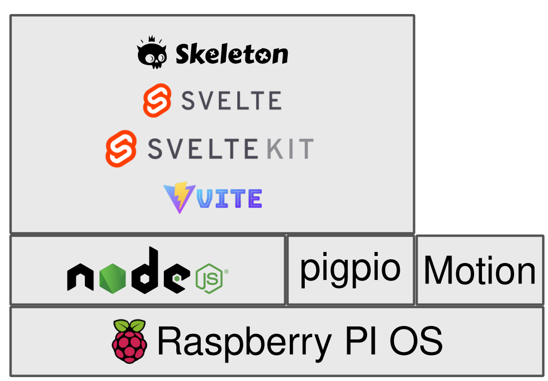


Figure 3. RaspCon software stack.

benefits is its developer-friendly nature resulting in minimal boilerplate code. Leveraging a component-based architecture that seamlessly and type-safely integrates server and client functionality, it facilitates modular development and an overall cohesive development experience. This not only streamlines the development process but also enhances scalability and maintainability. Furthermore, SvelteKit's filesystem routing simplifies the navigation structure of our application, enhancing code clarity and reducing complexity. The framework's built-in support for automatic Server-Side Rendering (SSR) ensures optimal performance and user experience.

The hardware configuration data exhibits a document-like structure and is of relatively low size and complexity. Consequently, the use of a traditional relational or even document-based database is unnecessary. Instead, we employ JSON structures as a lightweight yet effective solution. An advantageous aspect of this approach is its facilitation of exporting configuration subsets, such as presets, as JSON files. This enables users to locally store and transfer their presets to other instances with similar hardware setups.

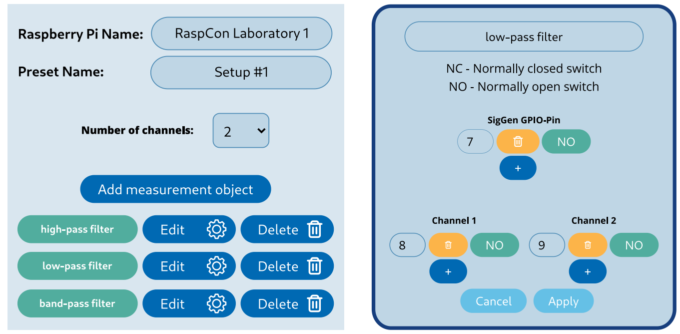
Although the system was designed as a web application, it required low-level access to the Raspberry Pi's GPIO pins to control both the electronic switch matrices and camera servo motors. This functionality was achieved by utilizing the C library pigpio, accessed through a JavaScript wrapper library. We constructed higher-level abstractions for all required hardware functions. While accessing the camera interface, transcoding video material and serving a live stream within JavaScript can be complex, we have found a straightforward solution with the software Motion. Utilizing this standard application enables us to effortlessly serve the video stream as MJPEG over HTTP, streamlining our implementation process.

V. Web User Interface

In the context of remote laboratory access, particularly for users situated beyond the confines of the university, the RaspCon application introduces a Web UI acting as a client interface through the integration with the LabCon RLS. This Web UI is designed to provide a seamless user experience across three key pages:

- Homepage: Introduces the RaspCon System, succinctly outlining its purpose and capabilities for remote experimentation, thereby orienting users to the system's functionality.
- Configuration Page: Reserved for administrators, this page offers a suite of tools for setting up and managing experimental configurations.
- Measurement Object Page: Enables students to engage in experiments by connecting measurement objects with measurement devices, enhanced by a live lab camera feed.

In the RaspCon application, the configuration page serves as a key interface for administrators, facilitating the setup of experimental environments at the start of each semester. This feature allows for the creation of a digital twin of the actual laboratory setup, incorporating the configuration of measurement objects and measurement devices to accurately



(a) Definition of a preset for connected measurement objects.

(b) Assignment of GPIO pins that control the switches for one measurement object.

Figure 4. Configuration page of an experimental setup.

reflect the physical lab arrangement (Figure 4a). The process includes simulating electronic switch matrices by assigning the GPIO pins controlling the individual switches and defining their initial states. This emulation helps to replicate the operation of individual switches, effectively merging the virtual and physical laboratory experiences (Figure 4b), thereby enhancing the digital twin's realism. Additionally, the configuration page offers comprehensive tools for managing experimental setups. Administrators can export and import configurations locally, and save and load them from the RaspCon server. This flexibility supports easy sharing and replication of setups across semesters, allowing for efficient organisation and retrieval of diverse experimental environments.

In the RaspCon system, the measurement object page is specifically designed for students to interact directly with laboratory hardware when conducting their assigned practicals during the semester, underpinned by configurations set by administrators before. This interface allows students to connect selected measurements objects with necessary measurement devices for a specific experiment, using a visual representation of the switch matrix displayed in table form (Figure 5). Through this interface, connections between measurement objects and measurement devices can be easily established or removed by clicking a button corresponding to the desired interaction, ensuring a straightforward and intuitive user experience. To maintain the integrity of experimental setups and prevent misuse, the system automatically removes an existing connection between a measurement object and a measurement device channel when a new measurement object is assigned to that channel. The measurement object page also integrates a live stream feature of the laboratory, enhancing the hands-on experiment with visual feedback. Students can manipulate the camera angle using arrow buttons, offering personalized views of the laboratory setup matching their experimental focus, or reset the camera to a standard position with a dedicated button (Figure 6). This live feed allows students to monitor the actual laboratory environment in real-time fostering interactive

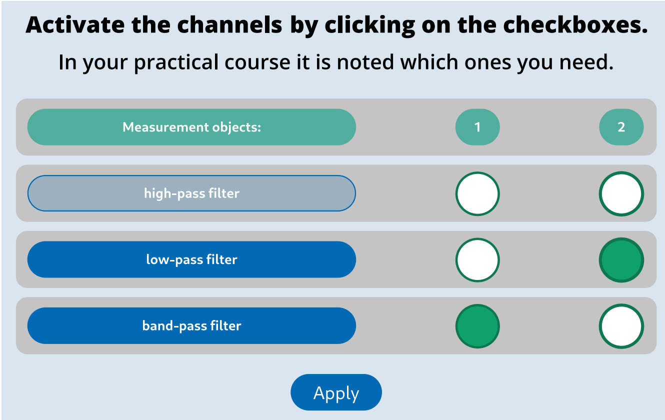


Figure 5. Example switch matrix table used to select active connections (green circles) between measurement objects and measurement device input channels.

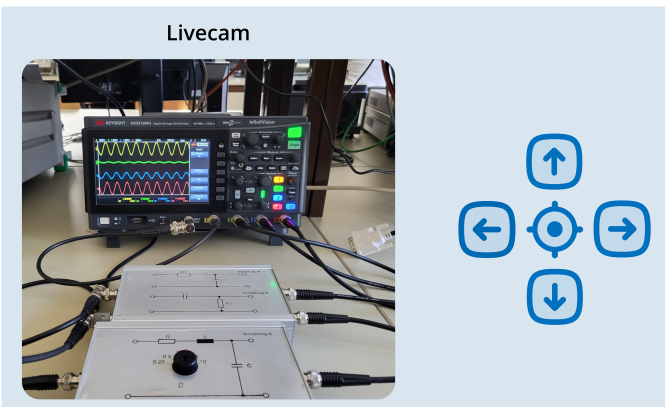


Figure 6. Integrated camera live stream.

learning and simulating physical presence in the laboratory, despite remote access.

The development of the Web UI for the RaspCon application was achieved through the utilization of Skeleton (Figure 3), a development toolkit tailored for integration with the Svelte library and enhanced by the capabilities of Tailwind CSS. Skeleton's provision of pre-configured components significantly streamlined the process of constructing consistent and contemporary user interfaces. A notable feature includes the implementation of a theme switcher, which greatly simplified the introduction of a dark mode. Moreover, the toolkit's predefined drawer and modal components were instrumental in organizing the pages' layout, ensuring a user-friendly and intuitive interface. To safeguard against unintentional modifications, e.g., of the experimental configurations currently in use, the Web UI incorporates confirmation dialog windows that prompt users for verification before proceeding with the deletion of settings.

VI. Quality Control And User Feedback

To maintain a high standard of quality for RaspCon, the web application underwent evaluation by its potential end-users, including students and administrators, on two occasions. The objective of these evaluations was to ensure that RaspCon offered an easy and intuitive user experience and to refine the application based on the feedback received.

For the efficient collection and analysis of feedback across all test groups, the web based tool AttrakDiff [10] was utilized. AttrakDiff is a questionnaire-based tool used to measure the perceived usability and aesthetic appeal of products and interfaces by assessing their pragmatic and hedonic qualities. It enabled the systematic gathering of responses through a structured questionnaire completed by each participant. The results of this evaluation are depicted in a portfolio format, as shown in Figure 7.

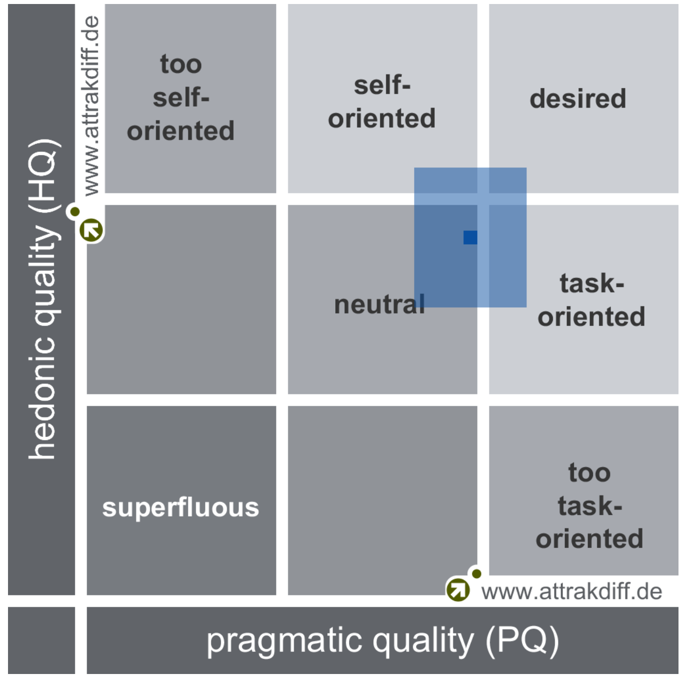


Figure 7. Portfolio format of RaspCon evaluation assessing hedonic and pragmatic quality.

The analysis reveals that RaspCon is perceived positively, demonstrating a tendency towards a favorable user experience characterized by both hedonic qualities (such as perceived stimulation and identification with the user) and pragmatic qualities (emphasizing usability and achievement of objectives).

VII. Conclusion and Future Work

The article introduces RaspCon, a scalable module designed to enhance remote laboratory experimentation within the LabCon RLS. It provides a web-based platform for administrators to configure and manage laboratory setups, allowing flexibility and interchangeability for educational needs. This system enables students to perform remote experiments, facilitating dynamic adjustments to experimental configurations as required. To further enrich the educational journey, the interaction capabilities will be extended and an AI agent is to be integrated providing dedicated support for the laboratory experiments.

REFERENCES

[1] P. Orduña, L. Rodriguez-Gil, J. Garcia-Zubia, I. Angulo, U. Hernandez, and E. Azcuenaga, "LabsLand: A sharing economy platform to promote educational remote laboratories maintainability, sustainability and adoption," in *2016 IEEE Frontiers in Education Conference (FIE)*, 2016, pp. 1–6.

[2] Keysight Technologies, *PathWave Lab Operations for Remote Learning, PW9112EDU*, 2021, retrieved: April, 2024. [Online]. Available: https://www.keysight.com/de/de/assets/3121-1016/data-sheets/PathWave-Lab-Operations-for-Remote-Learning-PW9112EDU.pdf.

[3] J. A. Ariza and C. N. Galvis, "RaspyControl Lab: A fully open-source and real-time remote laboratory for education in automatic control systems using Raspberry Pi and Python," *HardwareX*, vol. 13, p. e00396, 2023.

[4] M. C. Niquin, N. C. Puelles, M. C. Roman, J. V. Araoz, V. Pimentel, and C. L. Pineda, "Low-Cost Remote Laboratory for Robotics Undergraduate Learning during Covid Context," in *2022 IEEE 2nd International Conference on Educational Technology (ICET)*, 2022, pp. 11–15.

[5] A. Fernández-Pacheco, S. Martin, and M. Castro, "Implementation of an Arduino Remote Laboratory with Raspberry Pi," in *2019 IEEE Global Engineering Education Conference (EDUCON)*, 2019, pp. 1415–1418.

[6] J. Grodotzki, T. R. Ortelt, and A. E. Tekkaya, "Remote and Virtual Labs for Engineering Education 4.0: Achievements of the ELLI project at the TU Dortmund University," *Procedia Manufacturing*, vol. 26, pp. 1349–1360, 2018, 46th SME North American Manufacturing Research Conference, NAMRC 46, Texas, USA.

[7] S. García, A. Gallardo, D. F. Larios, E. Personal, J. M. Mora-Merchán, and A. Parejo, "Remote Lab Access: A Powerful Tool Beyond the Pandemic," in *2022 Congreso de Tecnología, Aprendizaje y Enseñanza de la Electrónica (XV Technologies Applied to Electronics Teaching Conference)*, 2022, pp. 1–5.

[8] M. Achilles *et al.*, "Enhancing Digital Learning: A User Management and Access System for Remote Laboratories," in *Proceedings of the Fifteenth International Conference on Mobile, Hybrid, and On-line Learning (eLmL 2023)*. Venice, Italy: IARIA, April 2023, pp. 44–45.

[9] J. Zhao, "Design, simulation, prototypical implemetation and test of a modul for internet-based switching between test objects in laboratory practicals," Bachelor's thesis, University of Applied Sciences Mittweida, Mittweida, Germany, 10 2023.

[10] M. Hassenzahl, M. Burmester, and F. Koller, *AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität (AttrakDiff: A Questionnaire for Measuring Perceived Hedonic and Pragmatic Quality)*. Wiesbaden: Vieweg+Teubner Verlag, 2003, pp. 187–196.

Digital Peer Grading in Group Learning: Empirical Insights and Best Practices

Eduard Klein
Business Department
Bern University of Applied Sciences (BUAS)
Bern, Switzerland
e-mail: eduard.klein@bfh.ch

Claus Noppeney
Business Department
Bern University of Applied Sciences (BUAS)
Bern, Switzerland
e-mail: claus.noppeney@bfh.ch

***Abstract*—The impact of digital peer grading on group learning dynamics in project-based courses in higher education is studied using the web-based tool Peer Grading Tool (PGT). Following an action research approach, this study, which was conducted over two years across 20 courses, examines the use of PGT for both grade influencing and reflective peer assessment. Empirical data collection from students and lecturers led to interesting findings. The results reveal nuanced insights into the contextual suitability of PGT, highlighting factors, such as group size, familiarity among members, heterogeneity, voluntariness, and robust feedback culture. The research also highlights potential challenges and emphasizes the importance of a pre-existing feedback culture to mitigate negative impacts. The findings contribute to the derivation of actionable recommendations and best practices for the implementation of peer grading campaigns in higher education.**

Keywords-peer grading; group learning; peer assessments; free riding; group work reflection.

I. INTRODUCTION

In practice-oriented higher education, group work is often employed to facilitate independent and problem-oriented learning, as it effectively prepares students for project-oriented and agile work environments [1]. Ideally, all team members contribute equally to the overall success of the project, albeit often in individually diverse ways [2][3].

Peer assessment and peer grading are both methods of evaluating a peer's work, but they differ in their purpose and process. Peer assessment is a formative approach that aims to help students plan their learning, identify strengths and weaknesses, and develop personal and professional skills [4]. It involves students providing feedback to their peers about their work. On the other hand, peer grading is a more evaluative process where students assign grades or scores to their peers' work [5].

Peer grading is also a widely adopted method for group members to reflect on collaboration and evaluate individual contributions to the overall outcome. Through feedback from peers, individual learning opportunities are unlocked. Peer grading also serves to counteract undesirable free riding, which contradicts fairness principles and undermines the motivational functions of performance assessments.

Resulting from peer grading are dual outcomes; firstly, it serves as a mechanism for reflective practices within group learning processes. Secondly, the outcomes serve as a basis for individualized grading decisions. This can be through a *reflective grading* approach (see Section II), wherein peer feedback impacts the group process without directly impacting individual grades. Alternatively, an *individual grading* approach, aligned with the *assessment of learning* concept [6], involves using peer grading results to deviate grades from the group result potentially.

This duality underscores the nuanced role of peer grading in shaping not only individual grades but also the broader learning experience within collaborative group contexts. By adopting this didactic approach, lecturers harness the potential of peer grading to promote collaborative learning, encourage self-reflection, and drive iterative improvements in both group processes and outcomes.

Peer grading also unlocks valuable learning opportunities while simultaneously avoiding unwanted social loafing. The mere expectation of negative feedback from peers triggers positive behavioral changes in learners [7]. Free riding poses a central challenge in group-oriented learning contexts [8], as performance-oriented students perceive group work with free riding as an overall frustrating experience [9].

The rest of this paper is organized as follows. Section II provides an overview of our research design. In Section III, the peer grading tool PGT is described. Section IV presents the results of accompanying research on how the use of digital peer grading in group learning and presents best practices. The conclusions in Section V close the article.

II. RESEARCH DESIGN

Following an action research approach, we study the use of digital peer grading and its implications using the web-based *Peer Grading Tool* (PGT) [10]. Our goal is to contribute to the improvement of teaching in higher education: How can digital peer grading enhance learning in groups for students? How can it improve our teaching and coaching of group work as lecturers?

Empirically, we report on data generated over 4 consecutive semesters (Q4/2021-Q2/2023) in 20 different courses using the peer grading tool. The data collection was conducted using both qualitative and quantitative methods (survey, focus groups) and involved students as well as lecturers.

Our research covers different types of assessment as well as of peer grading. Depending on the didactic setting, there are several types of assessments. *Assessment of learning* focuses on measuring students' knowledge and skills, often through tests and exams [11]. In contrast, *assessment as learning* involves students in the assessment process, promoting self-regulation and decision-making [12]. This approach aligns with the concept of *assessment for learning*, which emphasizes using assessment to improve learning and provide feedback [13][14].

Two types of peer grading, *reflective* and *individual*, were used in this study. The *individual* approach is employed at the end of group work and falls into the category of assessment of learning. The *reflective* approach is also applied during group activities and falls into the category of assessment as learning or for learning.

In this study, the individual approach was also used to individualize group marks. To clarify, the term "marks" refers to the deliverables assessed by the lecturer, while the term "grade" refers to the point scale of the peer assessment criteria. For each group member, the deviation from the group average of peer grading is calculated (see Section IV), and thresholds for these variations are established. If a group member's deviation from peer grading surpasses an upper (or lower) threshold, defined for a course, an improvement (or deterioration) of the individual mark compared to the group mark can be made. This method allows for the individualization of group marks.

The surveys used structured as well as unstructured questions and covered a broad range of relevant topics (e.g. prior experiences of groups learning; attitude towards group learning, group dynamics or work attitude, etc.). In particular, the survey focused on the respondent's experience of peer grading. The following list of statements serves as a sample item. The respondents were asked to indicate their approval of the following statements on a 5-point scale (1 = "completely disagree"; 5 = "completely agree"):

(1) Peer grading has made free riding more difficult.
(2) Work was faked in the group to receive a good grade in peer grading.
(3) Peer grading made it possible to reflect on one's own role or group behavior.
(4) Peer grading increased the pressure within the group.
(5) In peer grading, students tried to evaluate actual performance.
(6) Peer grading makes the module assessment fairer.
(7) In addition to actual performance, students also considered irrelevant criteria (e.g. sympathy) when grading their peers.
(8) Peer grading enables students to work better in upcoming group work.
(9) Peer grading makes it easier to recognize the performance of all group members.
(10) Peer grading strengthens the feeling of trust in the team.

Section IV discusses the outcomes of the evaluations in more detail.

III. THE PEER GRADING TOOL PGT

The search for suitable peer grading tools tailored to our teaching contexts led to the evaluation of several systems, which we examined based on our specific requirements. It is noticeable that comparable peer grading tools are predominantly available through commercial licensing models, exemplified by Purdue University's CATME [15], or have limitations such as dependency on participants having a Google account, as with a peer grading tool integrated into Google Spreadsheets [16], which contradicts intended principles of flexibility, openness, and free accessibility.

The Moodle activity *Workshop* has similarities with the planned peer grading scenario. However, the analysis revealed that key features, encompassing functionality, granularity, flexibility, and customization, are not available in the Moodle activity *Workshop*. Therefore, the decision was made to implement the PGT tool as an in-house development and use Angular technology to create a publicly accessible web-based application.

We considered the following key requirements:

(1) Each group member can grade each other group member (peer assessment) and also herself (self-assessment) according to defined criteria based on a numeric scale.
(2) The criteria (names) and the numeric grade scale are definable by the lecturer during the creation of a peer grading campaign. The number of criteria (currently up to 10) and the criterias' weights are definable.
(3) The criteria are not necessarily focused on some artifact or group product. Instead, they could also focus on the group process or project management aspects.
(4) In case some or all criteria focus on an artifact, there is no upload required (as forced by some other tools). Instead, artifacts are communicated or shared outside the tool, e.g., via a learning platform.
(5) The criteria catalog contains a description of each criterion's grade to enable objective and equal grading for all groups. By decision, it cannot be imported into the tool to avoid the complex handling of the possible variety of semi-structured documents. Instead, it must be provided outside the system.
(6) Participants' data, such as name, e-mail address, etc., are provided as CSV data for bulk import. The tool provides an appropriate template.
(7) Privacy and data protection are enforced: after a peer grading campaign, each student receives system-generated feedback containing the deviation of his self-assessment grade from the peer assessment grade. The peer assessment grade for student x is calculated as the average of his peers' grades concerning student x.
(8) To facilitate tool use, students don't have to register, but instead receive a system mail with a personalized link to his or her grading form (see Figure 1).
(9) Lecturers must register since they carry out a workflow and must come back into their tool session from time to time.

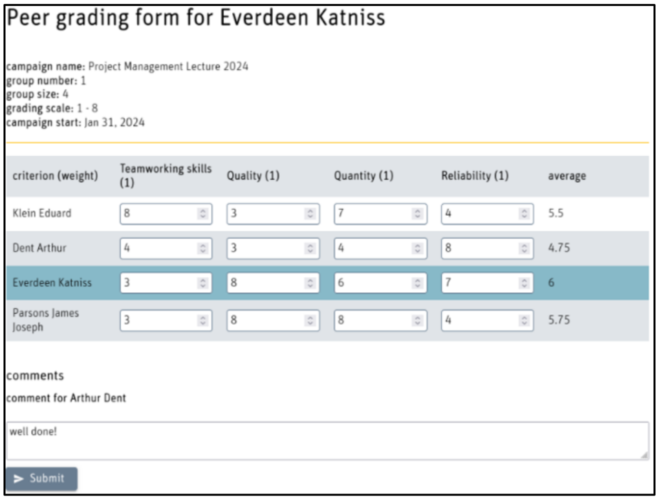


Figure 1. Peer grading form for Katniss Everdeen.

The creation of a peer grading campaign by the lecturer is shown in Figure 2, demonstrating flexibility in creating and editing criteria.

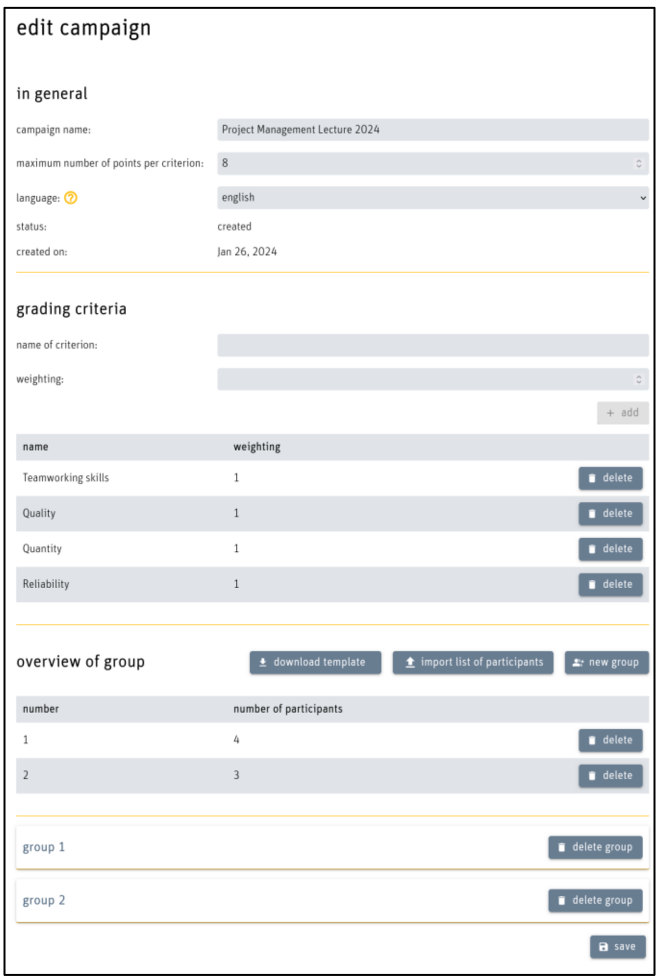


Figure 2. Creating a peer grading campaign.

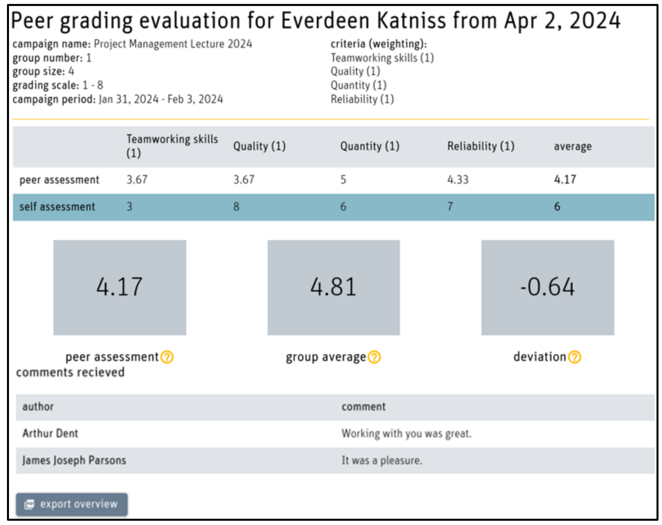


Figure 3. Peer grading result for Katniss Everdeen.

Upon completion of a peer grading campaign, all participants and the lecturer receive evaluations via system-generated e-mails. Each student receives anonymized feedback (see Figure 3) concerning the grading from his peers, and optional textual feedback as comments. The lecturer receives each other's gradings and the detailed gradings from and to each other's students (see Figure 4). Figure 3 shows the peer grading results for Katniss Everdeen, presenting her peers' grading per criterion, her self-grading, and (the big numbers from left to right) the average of her peers' gradings, the group average (=average of all peer gradings), and her derivation from the group average.

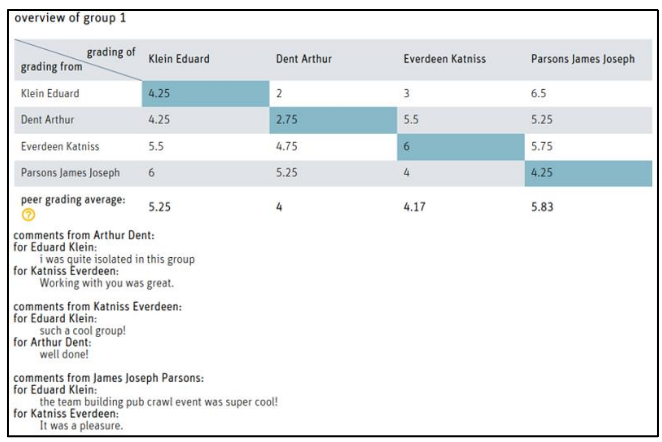


Figure 4. Lecturer's view of how group members graded each other.

The lecturer has control over the peer grading process at any time (see Figure 5). The functionality comprises editing campaign details, reminding defaulting students, closing campaigns, and generating detailed result views.

The software is available as open-source under MIT license at https://github.com/digital-sustainability-lab/peer-grading-tool-mirror.

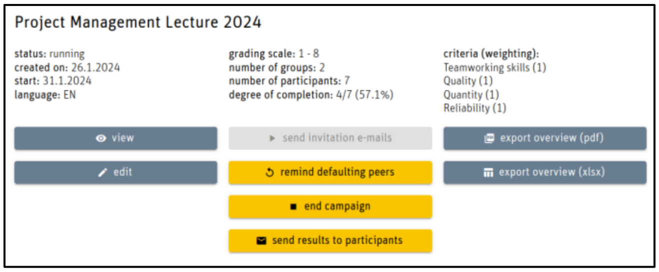


Figure 5. The lecturer has full control over the peer grading process.

Concerning the criteria catalog, for each criterion, the meaning of a grade must be defined. Guidelines for developing appropriate criteria are outside the focus of this paper. Instead, we list the grades' meanings for the criteria *ability to work in a team* for one of our study modules which has been proven appropriate for many years:

8/7: Exceptional contribution to the group process; "doer"; proactively advances project goals; is extremely committed; "sacrifices" himself/herself for project success; accomplishes much more than expected; opinion leader; high communication skills.

6/5: Committed member, contribution not too high, not too low; participates in the group process with an average sense of responsibility; has an integrating effect in conflicts.

4/3: Provides contributions on demand; neutral attitude towards project success; does not make motivational advances when there are "sags" in the team.

2/1: Unproductive to counterproductive attitude; does not contribute to the success of the project or hinders/prevents its progress. Destructive charisma.

In our study modules using peer grading, the criteria catalog is provided on the learning platform and discussed before the start of a peer grading campaign.

IV. RESULTS AND EXPERIENCES FROM PEER GRADING

The section covers three levels of results. Initially, the experience with individual grading is considered where the result of peer grading might influence the individual marks of the students. Second, results from the accompanying research are discussed, and finally, best practices are discussed.

A. *Individual Grading*

Figure 6 shows the results of a typical peer grading in a course. The x-axis represents the deviation of each student's grade compared to the group grade (which is the average of all peer grades in the group).

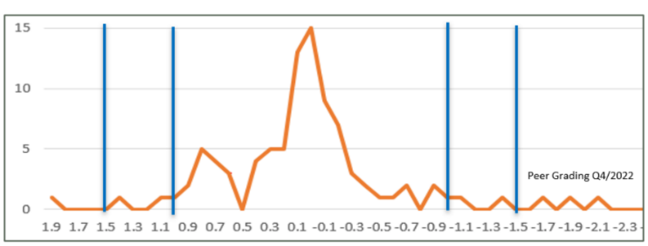


Figure 6. Results from one peer grading campaign. The x-axis shows deviations from the group grade, the y-axis shows the respective number of students.

The y-axis shows the number of students per deviation step. The blue lines define (arbitrary) thresholds at which the deviation leads to an individualized mark. The distribution shows that less than 10% of the students have a high or even extreme deviation. An analysis of the grades reveals that no discrimination based on gender or other diversity parameters can be found. Also, when comparing different campaigns, it is noticeable that the distribution is not influenced by the type of peer grading (i.e., reflective, or individual grading).

Occasionally, group-internal agreements such as identical peer gradings can be observed in the data. Although this may seem like a subversive practice that undermines the didactic concept, it is perfectly acceptable, as the individualization of marks should be limited to rather extreme cases. However, there is the scenario that multiple free riders in one group can potentially coordinate their grading activities and even downgrade potential outperformers.

Regardless of these group dynamic effects, our conclusion and a strict recommendation are: In the case of individual grading, the results of peer grading should not automatically lead to a change in the individual mark (see Figure 6). Instead, high deviations in peer grading should be compared with the lecturer's impression of the group and trigger a specific debriefing session with the group. This aligns with the fact that students sometimes tend to rate their peers' performance differently than their lecturer does and that peer evaluations may not always accurately reflect performance [17].

B. *Experience with Peer Grading*

The accompanying research on the experiences of students and lecturers (including surveys and focus groups) provides important insights.

In the students' feedback, several critical observations emerged in the free text fields. Firstly, there were concerns about the perceived fairness of the assessments. Some students expressed dissatisfaction with assessments that they felt were unfair and highlighted the need for transparency and consistency in grading.

In addition, students pointed out the influence of social factors such as likability or familiarity among fellow students, which they felt could have a subjective impact on the evaluation of fellow students. Finally, students shared their concern that peer grading could draw too much attention to the assessment tasks and thus undermine the learning experience.

However, the few consistently critical opinions expressed in the answers to the open questions are not confirmed by the quantitative results. The following statements receive high and very high approval rates.

- In peer grading, students tried to evaluate actual performance.
- Peer grading has made free riding more difficult.
- Peer grading makes the module assessment fairer.

The following statements continue to receive moderate approval:

- Peer grading made it possible to reflect on one's own role or group behavior.
- Peer grading increased the pressure within the group.

- In addition to actual performance, students also considered irrelevant criteria (e.g., sympathy) when grading their peers.
- Work was faked in the group to receive a good grade in peer grading.

Several of these statements have been derived from research findings, with our focus centered on the students' emphasis placed on statements related to peer grading. As per [17], peer grading can indeed make free riding more difficult, as it can lead to more accurate and reliable assessments. Furthermore, peer grading has been shown to facilitate reflection on one's role in group work [18].

Accordingly, important goals of peer grading can be achieved: Making free riding more difficult facilitates and promotes group learning. In addition, the final assessment is perceived as fair. Since our approach is not an experiment but action research, it is difficult to capture or even measure the effects of peer grading. Therefore, we asked students to compare their experiences with group work in different courses.

The analysis of the responses from the follow-up surveys among students yielded interesting findings:

- On average, peer-graded group work received comparable or better grades than other group work in all courses observed.
- Students acknowledge that peer grading reduces free riding, but this effect also leads to divergent opinions. It is crucial to recognize that, similar to an echo chamber, individual participants reinforce negative feelings and can thus have an unfavorable influence on the formation of opinions in the group.
- A small proportion of around 10-25% are against the continued use of peer grading.
- Conversely, around 75-90% are in favor of the continued use of peer grading. The distribution of the surveyed frequencies (rarely, occasionally, often, always) varies considerably depending on the course.
- A small proportion of 7-13% of students are in favor of the use of peer grading in all group work.

A correlation between rather negative statements regarding the use of peer grading and weak peer grading results of individual persons could not be investigated, as the surveys were conducted anonymously. For the same reason, it is not possible to determine which group of people took part in the surveys about peer grading and in what proportions (response rate approx. 30% across all courses). In the debriefing discussions, however, it became obvious that opinion leaders often come from a group of people who either contributed little to the group work and/or had poor feedback in the peer grading.

C. Best Practices

There are several contextual factors to consider when introducing digital peer grading. Peer grading seems particularly appropriate when student groups tend to be larger. This practical insight is in line with the fact that once a group is formed, group size influences the decisions of members to contribute to the group's public good. In a small group members can easily notice if a member does not equally contribute to the group's efforts. However, as group size increases, free riding becomes more probable [19]. In this case, digital peer grading can compensate for the decreasing noticeability and perceptibility.

Another contextual factor to be considered is the familiarity of the team members. In scenarios where group members have limited knowledge of each other, peer grading can provide a fairer and more impartial assessment framework.

In addition, the heterogeneous composition of groups can benefit from peer grading as it allows different perspectives to be considered in the assessment. In cases where group formation is less voluntary, such as assigned or structured groups, peer grading provides a method of impartial assessment.

In educational settings where intense collaboration and self-organization are expected, peer grading also fosters a sense of ownership and responsibility among students.

Finally, a developed feedback culture helps when digital peer grading is used. In this case, digital peer grading offers new learning opportunities and serves as an acceptable method to reflect on the collaboration. If the feedback culture is not part of the didactic setting the use of digital peer grading might provoke dysfunctional group behavior and undermine group learning.

In summary, digital peer grading turns out to be suitable if the following characteristics are present:

- larger groups (>3 members)
- less familiar
- more heterogeneous
- less voluntary composition
- intense collaboration, self-organized
- higher feedback culture

Both the survey results and the analysis of the evaluations prove to be quite stable. They hardly deviate from the experiences documented in this section, not even for peer gradings in new classes.

V. CONCLUSION AND FUTURE WORK

Awarding identical marks in group work to all members of a group is only considered fair if all group members contribute to the group result in roughly the same way. In this study, peer grading is used on the one hand as a way of nuancing the group mark and on the other hand to reflect on the group work. To minimize the workload for lecturers and students concerning peer grading, a digital peer grading tool is used that automatically compares self-assessment and peer assessment and calculates individual deviations of group members from the group average.

The experiences and best practices from Section IV have so far been repeatedly observed in new peer grading courses, which is a sign of stability for best practices. A larger study is planned for the future, which will carry out extended evaluations and analyses about a possible segmentation of courses into groups of courses with similar characteristics (study program, group work skills etc.). This will be possible because new courses from other disciplines will be added as the PGT tool becomes more widespread. Specific experiences and best practices may emerge, differentiated by segment.

Extensions are also planned concerning the (multilingual) peer grading tool. PGT currently runs as a single server instance including a database in which all user data is stored. For data protection and performance reasons, use with third-party hosted containers and the possibility for own branding will also be offered, including analytics for an overview of the use of installed instances.

ACKNOWLEDGMENT

The PGT tool was developed with funding from the BUAS Virtual Academy (Bern University of Applied Sciences). The study was conducted in the context of the DigiPeG Project [20] which has funding from the BeLEARN association which is an initiative of the Canton of Berne.

REFERENCES

[1] A. S. Burke, “Group Work: How to Use Groups Effectively,” *The Journal of Effective Teaching*, vol. 11, pp. 87–95, 2011, [Online]. Available: https://api.semanticscholar.org/CorpusID:265889597

[2] D. B. Kaufman, R. M. Felder, and H. Fuller, “Peer ratings in cooperative learning teams,” *ASEE Annual Conference Proceedings*, no. June, pp. 3931–3942, 1999, doi: 10.1.1.193.23.

[3] S. Brutus and M. Donia, “Improving the effectiveness of students in groups with a centralized peer evaluation system,” *Academy of Management Learning and Education*, vol. 9, no. 4, pp. 652–662, 2010, doi: 10.5465/AMLE.2010.56659882.

[4] K. J. Topping, “Peer Assessment,” *Theory Pract*, vol. 48, no. 1, pp. 20–27, 2009.

[5] H. Li *et al.*, “Peer assessment in the digital age: a meta-analysis comparing peer and teacher ratings,” *Assess Eval High Educ*, vol. 41, no. 2, pp. 245–264, 2016, doi: 10.1080/02602938.2014.999746.

[6] D. Wiliam, “What is assessment for learning?” *Studies in Educational Evaluation*, vol. 37, no. 1, pp. 3–14, 2011, doi: https://doi.org/10.1016/j.stueduc.2011.03.001.

[7] C. M. Brooks and J. L. Ammons, “Free Riding in Group Projects and the Effects of Timing, Frequency, and Specificity of Criteria in Peer Assessments,” *Journal of Education for Business*, vol. 78, no. 5, pp. 268–272, 2003, doi: 10.1080/08832320309598613.

[8] P. Aggarwal and C. L. O’Brien, “Social loafing on group projects: Structural antecedents and effect on student satisfaction,” *Journal of Marketing Education*, vol. 30, no. 3, pp. 255–264, 2008, doi: 10.1177/0273475308322283.

[9] D. Hall and S. Buzwell, “The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution,” *Active Learning in Higher Education*, vol. 14, no. 1, pp. 37–49, 2013, doi: 10.1177/1469787412467123.

[10] E. Klein, “Peer Grading Tool - PGT,” Website. Accessed: Apr. 15, 2024. [Online]. Available: http://www.bfh.ch/peer-grading-tool

[11] O. Novogen, B. Nina, and E. A. Railean, “Assessment as Learning,” *Assessment, Testing, and Measurement Strategies in Global Higher Education*, pp. 238–255, 2020, [Online]. Available: https://api.semanticscholar.org/CorpusID:213527226

[12] R. Dann, “Assessment as learning: blurring the boundaries of assessment and learning for theory, policy and practice,” *Assess Educ*, vol. 21, pp. 149–166, 2014, [Online]. Available: https://api.semanticscholar.org/CorpusID:62881136

[13] S. Rahman and A. A. Majumder, “Is it assessment of learning or assessment for learning,” *Southeast Asia Journal of Public Health*, vol. 4, pp. 72–74, 2015, [Online]. Available: https://api.semanticscholar.org/CorpusID:73149255

[14] S. Chappuis and R. J. Stiggins, “Classroom Assessment for Learning.,” *Educational Leadership*, vol. 60, pp. 40–43, 2002, [Online]. Available: https://api.semanticscholar.org/CorpusID:142695229

[15] A. Loignon *et al.*, “Facilitating Peer Evaluation in Team Contexts: The Impact of Frame-of-Reference Rater Training,” *Academy of Management Learning & Education*, vol. 16, no. 4, pp. 562–578, 2017.

[16] S. Stolz, “Peer-Grading Tool.” Accessed: Apr. 15, 2024. [Online]. Available: https://gitlab.com/st.stolz/Peer-Grading-Tool

[17] P. Sadler and E. Good, “The Impact of Self- and Peer-Grading on Student Learning,” *Educational Assessment*, vol. 11, no. 1, pp. 1–31, Feb. 2006, doi: 10.1207/s15326977ea1101_1.

[18] D. Nicol, A. Thomson, and C. Breslin, “Rethinking feedback practices in higher education: a peer review perspective,” *Assess Eval High Educ*, vol. 39, no. 1, pp. 102–122, Jan. 2014, doi: 10.1080/02602938.2013.795518.

[19] R. Albanese and D. D. Van Fleet, “Rational Behavior in Groups: The Free-Riding Tendency,” *Academy of Management Review*, vol. 10, no. 2, pp. 244–255, 1985, doi: 10.5465/amr.1985.4278118.

[20] E. Klein and C. Noppeney, “DigiPeG – Digitales Peer Grading für lernfördernde Gruppenarbeiten,” Website. Accessed: Apr. 15, 2024. [Online]. Available: https://belearn.swiss/en/projekt/digipeg-digitales-peer-grading-fuer-lernfoerdernde-gruppenarbeiten/