Adaptation, automated feedback, engagement, and effectiveness in learning data science: a summary of five years of research

Pierpaolo Vittorini*,†, Ivan Letteri† and Tamsir Jobe†

University of L'Aquila, P.le S. Tommasi 1, 67100 Coppito, L'Aquila (Italy)

Abstract

Data science has become an essential subject in higher education. Learning data science is a complex process for students. It requires integrating and applying statistical and computational principles. Adopting teaching strategies and tools that may support them is a current and relevant challenge. In our research, we designed and developed an adaptive tool providing adaptive formative feedback to students in data science courses. The tool uses artificial intelligence technologies to tailor the exercises and the returned feedback to students. The paper summarizes five years of research. We started with "good-old-style" AI technologies (i.e., static code analysis, embeddings, classifiers, regressors) and are now experimenting with LLMs to support feedback generation. We collected students' learning outcomes, engagement, and comments over the years. We exploited the comments to revise and improve the tool to fulfill the students' needs. With this methodology, we measured increased learning outcomes and high engagement (with an interesting exception when using LLM feedback). The paper ends by discussing the findings and summarizing the lessons learned.

Keywords

AI, Generative AI, Technology-Enhanced Learning, Data science

1. Introduction

Data science has become a significant field within higher education in recent years. Mastering the subject poses complex challenges for students because it demands the assimilation, integration, and application of statistical and computational principles (e.g., data collection, data analysis, inference, algorithms, machine learning) [1]. In addition, the literature shows that students who attend data science courses have varying preparation, culture, and scholastic backgrounds [2]. Deploying effective teaching strategies and tools to aid data science students is, therefore, a relevant and ongoing challenge. At the University of L'Aquila, in the Department of Life, Health and Environmental Sciences, several degrees contain a course about the fundamentals of data science. These courses are attended by students and professionals in the health sector (e.g., diagnostic technicians) who need to learn how to organize datasets, apply descriptive and inferential statistics, and create predictive models. These students have the varying scholastic backgrounds reported in the literature, e.g., from young, motivated students with a good knowledge of mathematics and computer science, to old professionals only seeking an additional qualification for career advancement. Moreover, our classes are large, with a total number of students of around 200 people (which will probably double within a couple of years). Therefore, we started a research project aiming to design and develop a system able to provide *effective*, *automated*, and *elaborated feedback* to students when studying the course topics. This paper summarizes five years of research in this context, in terms of both the methodologies (formative assessment, scaffolding, feedback), technologies (automated adaptive feedback built on AI and Generative AI research), and achieved results (engagement, effectiveness) in teaching and learning data science.

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy

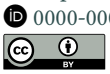
*Corresponding author.

†These authors contributed equally.

pierpaolo.vittorini@univaq.it (P. Vittorini); ivan.letteri@univaq.it (I. Letteri); jtamsir7@gmail.com (T. Jobe)

https://vittorini.univaq.it/ (P. Vittorini); https://ivanletteri.it/ (I. Letteri)

0000-0002-6975-8958 (P. Vittorini); 0000-0002-3843-386X (I. Letteri); 0009-0001-6653-8239 (T. Jobe)



2. Theoretical Framework

We grounded our research into a theoretical framework based on (i) formative assessment, (ii) scaffolding, and (iii) elaborated feedback. Formative assessment was initially introduced as a technique to provide feedback essential for analyzing learning developments [3]. It transitioned from merely being an "interim" evaluation tool for educators and learners to an ongoing process offering insights into student understanding, helping to adapt teaching approaches [4, 5]. Formative assessment is closely linked to scaffolding [6, 7, 8, 9], both aiming to guide learners from their current knowledge to the skills they can develop next, i.e., in the zone of proximal development [10]. Yet, while formative assessment adjusts teaching based on insights into the learner's comprehension, scaffolding offers support through hints and motivation. Feedback is a crucial point for both formative assessment and scaffolding [11], in particular, to support the building of new knowledge based on the existing understanding of the subject throughout the whole learning process. Three different types of feedback exist:

- KR : *Knowledge of Results* : it returns if the solution is right or wrong;
- KCR : *Knowledge of Correct Result* : if the solution is wrong, returns the correct solution;
- EF : *Elaborated Feedback* : if the solution is wrong, provides clues and hints useful to solve the exercise.

The meta-analysis discussed in [12] highlights that using computer-based formative assessment systems results in an average effect size of 0.28 on learning outcomes. The meta-analysis from Doo et al. [13] reports different effect sizes (depending on the specific type of scaffolding) ranging from 0.13 to 0.38. A recent meta-analysis [14] indicates that EF has a greater effect size (0.49) compared to KR (0.05) or KCR (0.32) on the learning outcomes, and that the effect size is adversely influenced by delayed feedback.

3. Applicative Scenario

During formative assessment, students can solve exercises that are helpful to better understand the course and to be prepared for the final exam. A sample exercise follows.

> *Consider dataset* `sbp`*. The dataset contains data on ten patients. For each patient, it includes the gender (variable* `gender`*), and the systolic blood pressure before (variable* `before`*) and after (variable* `after`*) taking an antihypertensive drug. Verify whether the difference in pressure after taking the drug between males and females is statistically significant, and comment on the test result. Submit as a solution a text that contains the list of R commands with the respective output and the comments given on the results.*

To solve the exercise correctly, a student must first identify the study design (it is a two-independent-samples design), then check the normality of the data (the samples are made up of only five patients each, so a normality test is mandatory), accordingly choose between a parametric or non-parametric test, and finally interpret the test result. A solution must be given in R, and must include all commands, the respective outputs, and the comments required for interpreting the results. A real solution provided by a student (incomplete, because it does not include the test for normality) is the following:

```
> t.test(sbp$after[sbp$gender=="m"], sbp$after[sbp$gender=="f"])

    Welch Two Sample t-test

data: sbp$after[sbp$gender == "m"] and sbp$after[sbp$gender == "f"]
t = 2.4771, df = 7.0303, p-value = 0.04224
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  1.544818 65.255182
```

```
sample estimates:
mean of x mean of y
     190.8      157.4

> # Given that the p-value is less than 0.05, I consider the difference
in pressure statistically significant and therefore generalizable to the
population.
```

A student submitting this solution will receive both schematic and detailed feedback, as follows. The system performs a static code analysis of the R commands and solutions (see §4.1), classifies the comments as right or wrong (see §4.1), calculates the distance of the solution with the correct one provided by the professor (see §4.1), and then produces a schematic feedback (see §4.2). The final result is shown on the left of Figure 1. In addition, the schematic feedback is exploited in a prompt that queries an LLM (see 4.4), in order to produce the detailed feedback shown on the right of Figure 1.

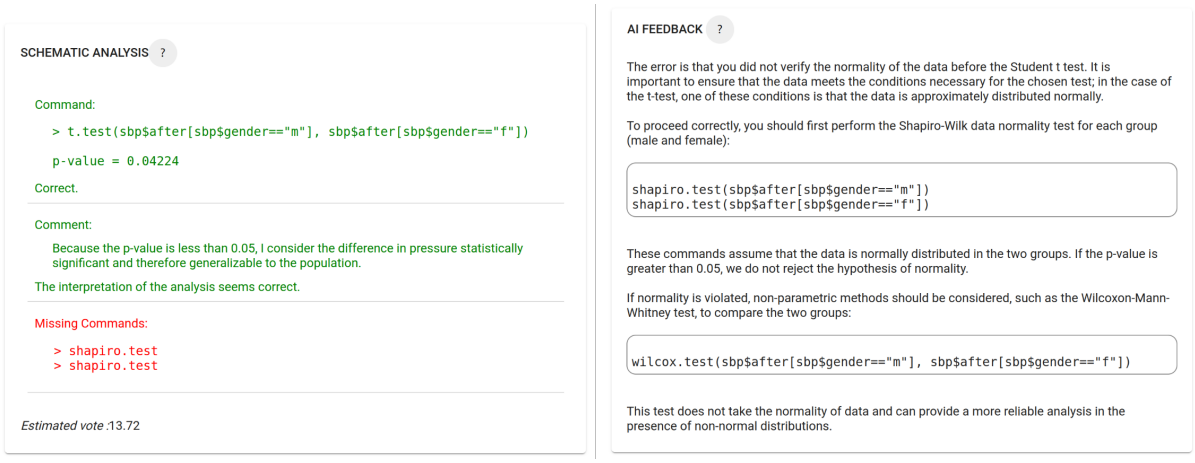


Figure 1: Schematic analysis (on the left) and detailed feedback (on the right)

Starting from the received feedback, the student can change the submitted solution, request a different exercise, or quit the platform. It is worth remarking that the exercise selected by the rDSA system, and how the schematic feedback is generated, depend on several adaptation criteria (see 4.3).

4. The Research

In the last five years, we pursued the following objectives. First, to develop a system that can analyse students' solutions and return an estimate of the final grade. Then, to improve the returned feedback in terms of elaborated feedback. Hence, introduce adaptation in selecting the exercises and in the feedback structure. Finally, to exploit LLMs to further improve the returned feedback.

4.1. AY 2020/21 – Develop a system that analyses students' solutions and returns an estimate of the final grade

The automated assessment of student programming assignments was first attempted in the sixties [15] and evolved during the years [16]. The automated analysis of open-ended answers has a long tradition in education [17], too. However, at the time of this research, no systems were available to analyze code for the R language, and no specific tools existed to classify sentences written in Italian that explain the result of a normality or hypothesis test. Therefore, we developed [18] a static code analysis tool for R (i.e., a tool that checks the commands and the respective output for errors without executing them) and a toolchain that uses FastText embeddings, hand-crafted features and a binary classifier for the comments (initially based on SVM [18], and then on MLP [19]). Moreover, to estimate the final grade, we introduced a "distance" that measures the similarity between the correct solution given by

the professor and a solution given by the student [18]. The main results we achieved are summarized in [18, 19]. In short: (i) we measured a good correlation between the manual grades and the estimated grades ($R^2 = 0.81$); (ii) we achieved an accuracy of 93% in classifying the comments as right or wrong; (iii) we observed a good user experience ($5.3/7$) and high engagement ($4.3/5$) of the students with the tool; (iv) students that did not use the system achieved an average grade of $21.7/30$, significantly lower than the others ($24.9/30$). As a summative assessment tool, it helped us to accelerate the correction process and reduce the mistakes, too.

4.2. AY 2021/22 – Improve the returned feedback

We improved the feedback [20] by introducing explanations on the possible reasons for the mistakes, which required: (i) introducing correct, wrong, and partially wrong commands; (ii) revising the distance definition; (iii) revising both the static code analyzer and the feedback generation system. This feedback is shown on the left of Figure 1. The main results we achieved are discussed in [20]. Students rated the general usefulness of the system with $4.2/5$, the clarity of the suggestions for the completely wrong commands with $3.5/5$, and the clarity of the suggestions for the partially wrong commands with $3.0/5$. The students' engagement remained high ($4.2/5$). Moreover, students who used the improved feedback achieved an average grade of $27.7/30$, whereas students who did not use the feedback achieved an average grade of $22.7/30$. Note that students of the previous year, who used the initial feedback, achieved an average grade of $24.9/30$.

4.3. AY 2022/23 – Introduce adaptation

As is known, adaptive learning systems personalize educational experiences by using data-driven techniques to tailor content, feedback, and pacing to individual learners. Research shows they improve engagement and outcomes, especially through technologies like intelligent tutoring systems, machine learning, and real-time learner modeling [21]. Accordingly, we introduced in the rDSA system an adaptation algorithm based on the analysis of the students' performances, using the Rasch model [22]. The main results we achieved are: good engagement ($4.0/5$), and average grades ($27.6/30$) similar to the previous year. Given these results, we improved the adaptation algorithm based on an ensemble regressor that can suggest to a student an exercise, which, if solved correctly, has the highest chances to improve his/her learning outcomes [23]. The novel adaptation algorithm halved the error in estimating the learning outcomes (from a MAPE of 39% with the Rasch model to 20% with the ensemble regressor).

4.4. AY 2023/24 and AY 2024/25 – Exploit LLMs to further improve the returned feedback

We introduced an LLM-based feedback. We tested different types of feedback, prompts, and models [24], until defining the detailed feedback (shown on the right of Figure 1). The main results we achieved are a reduced focused attention ($3.5/5$ compared with $3.8/5$ measured in the previous year). This indicates an increased cognitive load, which may be caused by the broader and more diverse information conveyed by the detailed feedback with respect to the schematic feedback. The average grades improved to $28.2/30$, and students who exercised more (i.e., > 90 exercises during formative assessment) achieved better grades than the others ($29.9/30$ vs $27.6/30$). Very recently [25], we carried out a study using LLMs GPT-4o and LLaMA 3.3-7B to grade student responses and generate constructive feedback based on student-submitted answers and corresponding reference solutions. We adopted a prompt engineering approach, iteratively refining prompts to optimize both the classification of short answers and the quality of feedback provided. The LLMs showed strong performance, comparable with that achieved with the ad-hoc toolchain (see 4.1), with GPT-4o achieving an accuracy of 93% and LLaMA 3.3-7B following closely at 92%. To assess the quality of the generated feedback, three domain experts evaluated the responses. The feedback was judged adequate in 76.7% of cases, with substantial inter-rater agreement as indicated by a Fleiss' Kappa score of 0.807.

5. Conclusions

Grounding our research in technology-enhanced learning (TEL) in a sound theoretical framework has shown its effectiveness. The results we achieved are manifold. We introduced a metric to measure the distance between solutions made up of commands (and their output) and comments. We developed a static code analyzer for R, an adaptive engine to select the "best" exercise for every student, two types of adaptive feedback (one based on "traditional" AI, another based on LLMs), tailored to data science for health professionals. Moreover, we improved the students' learning outcomes and achieved good engagement levels, therefore, adding further pieces of evidence about the pivotal importance of feedback during formative assessment to improve students' learning outcomes. At the same time, we highlighted possible drawbacks of EF generated by LLMs (i.e., cognitive load).

Even if the results focus on data science courses, the methodological choices (i.e., the theoretical framework, how to calculate the distance) are independent of the specific domain. In addition, the implemented tools have corresponding solutions for different languages/domains. For instance, many static code analysis tools exist for other programming languages (e.g., Cppcheck for C, Pylint for Python). Moreover, as already explored in §4.4, one can use LLMs instead of an ad-hoc toolchain for analyzing comments.

As future work, we are already working on using RAG/CAG to improve the quality of the returned feedback, whose preliminary results appear very promising. As a long-term implementation, we defined a plan to implement a module in the rDSA system for the semi-automatic generation of exercises with a structure, content, expected difficulty, and correct solution defined by the professor.

Declaration on Generative AI

During the preparation of this work, the authors used *Grammarly* for *grammar and spelling checks*. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. Buckingham Shum, M. Hawksey, R. S. Baker, N. Jeffery, J. T. Behrens, R. Pea, Educational data scientists: A scarce breed, ACM International Conference Proceeding Series (2013) 278–281. doi:10.1145/2460296.2460355.

[2] C. Brooks, R. M. Quintana, H. Choi, C. Quintana, T. NeCamp, J. Gardner, Towards Culturally Relevant Personalization at Scale: Experiments with Data Science Learners, International Journal of Artificial Intelligence in Education 31 (2021) 516–537. doi:10.1007/S40593-021-00262-2.

[3] M. Scriven, The Methodology of Evaluation, in: R. Tyler, R. Gagné, M. Scriven (Eds.), Perspectives of Curriculum Evaluation, AERA Monograph Series on Curriculum Evaluation, volume 1, Rand McNally, Chicago, 1967, pp. 39–83.

[4] R. E. Bennett, Formative assessment: a critical review, Assessment in Education: Principles, Policy & Practice 18 (2011) 5–25. doi:10.1080/0969594X.2010.513678.

[5] P. Black, D. Wiliam, Inside the Black Box: Raising Standards through Classroom Assessment, Phi Delta Kappan 92 (2010) 81–90. doi:10.1177/003172171009200119.

[6] D. Wood, J. S. Bruner, G. Ross, The Role of Tutoring in Problem Solving, Journal of Child Psychology and Psychiatry 17 (1976) 89–100. doi:10.1111/J.1469-7610.1976.TB00381.X.

[7] I. Clark, Formative Assessment: Assessment Is for Self-regulated Learning, Educational Psychology Review 24 (2012) 205–249. doi:10.1007/S10648-011-9191-6/.

[8] L. A. Shepard, Linking Formative Assessment to Scaffolding, Educational Leadership 63 (2005) 66–70.

[9] S. M. Kruiper, M. J. Leenknecht, B. Slof, Using scaffolding strategies to improve formative as-

sessment practice in higher education, Assessment & Evaluation in Higher Education 47 (2022) 458–476. doi:10.1080/02602938.2021.1927981.
[10] L. S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes, Harvard University Press, 1978. doi:10.2307/J.CTVJF9VZ4.
[11] B. Frank, N. Simper, J. Kaupp, Formative feedback and scaffolding for developing complex problem solving and modelling outcomes, European Journal of Engineering Education 43 (2018) 552–568. doi:10.1080/03043797.2017.1299692.
[12] N. Kingston, B. Nash, Formative Assessment: A Meta-Analysis and a Call for Research, Educational Measurement: Issues and Practice 30 (2011) 28–37. doi:10.1111/J.1745-3992.2011.00220.X.
[13] M. Y. Doo, C. J. Bonk, H. Heo, A Meta-Analysis of Scaffolding Effects in Online Learning in Higher Education, The International Review of Research in Open and Distributed Learning 21 (2020) 60–80. doi:10.19173/IRRODL.V21I3.4638.
[14] U. Mertens, B. Finn, M. A. Lindner, Effects of Computer-Based Feedback on Lower and Higher-Order Learning Outcomes: A Network Meta-Analysis, Journal of Educational Psychology 114 (2022) 1743–1772. doi:10.1037/EDU0000764.
[15] J. Hollingsworth, Automatic graders for programming classes, Communications of the ACM 3 (1960) 528–529. doi:10.1145/367415.367422.
[16] D. M. Souza, K. R. Felizardo, E. F. Barbosa, A Systematic Literature Review of Assessment Tools for Programming Assignments, in: 2016 IEEE 29th International Conference on Software Engineering Education and Training (CSEET), IEEE, 2016, pp. 147–156. doi:10.1109/CSEET.2016.48.
[17] S. Burrows, I. Gurevych, B. Stein, The eras and trends of automatic short answer grading, International Journal of Artificial Intelligence in Education 25 (2015) 60–117. doi:10.1007/s40593-014-0026-8.
[18] P. Vittorini, S. Menini, S. Tonelli, An AI-Based System for Formative and Summative Assessment in Data Science Courses, International Journal of Artificial Intelligence in Education (2020) 1–27. doi:10.1007/s40593-020-00230-2.
[19] A. M. Angelone, A. Galassi, P. Vittorini, Improved Automated Classification of Sentences in Data Science Exercises, in: F. De la Prieta, R. Gennari, M. Temperini, T. Di Mascio, P. Vittorini, Z. Kubincova, E. Popescu, D. R. Carneiro, L. Lancia, A. Addone (Eds.), Methodologies and Intelligent Systems for Technology Enhanced Learning, 11th International Conference, Springer, Cham, 2021, pp. 12–21. doi:10.1007/978-3-030-86618-1_2.
[20] P. Vittorini, A. Galassi, rDSA : an intelligent tool for data science assignments, Multimedia Tools and Applications 82 (2022) 12879–12905. doi:10.1007/s11042-022-14053-x.
[21] I. Gligorea, M. Cioca, R. Oancea, A. T. Gorski, H. Gorski, P. Tudorache, Adaptive Learning Using Artificial Intelligence in e-Learning: A Literature Review, Education Sciences 2023, Vol. 13, Page 1216 13 (2023) 1216. doi:10.3390/EDUCSCI13121216.
[22] G. Rasch, Probabilistic Models for Some Intelligence and Attainment Tests, Danmarks Paedagogiske Institut, 1960.
[23] P. Vittorini, A Report on the Use of the rDSA Tool for Formative and Summative Assessment, Lecture Notes in Networks and Systems 538 LNNS (2023) 23–32. doi:10.1007/978-3-031-20257-5_3.
[24] I. Letteri, P. Vittorini, Exploring the Impact of LLM-Generated Feedback: Evaluation from Professors and Students in Data Science Courses, in: C. Herodotou, S. Papavlasopoulou, C. Santos, M. Milrad, N. Otero, P. Vittorini, R. Gennari, T. Di Mascio, M. Temperini, F. De la Prieta (Eds.), Methodologies and Intelligent Systems for Technology Enhanced Learning, 14th International Conference, Springer Nature, 2024, pp. 11–20. doi:10.1007/978-3-031-73538-7_2.
[25] V. Cofini, T. Jobe, I. Letteri, P. Vittorini, Preliminary evaluation of an LLM-based system for grading and providing feedback on short-text answers in data science exercises, in: Methodologies and Intelligent Systems for Technology Enhanced Learning, 15th International Conference, Springer, Lille, 2025.